



INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

MHRA AI Airlock Phase 2: Seven AlaMD Evidence Case Studies

Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-10-13 · mhra ai airlock phase 2 · ai amd · ai medical devices · regulatory sandbox · clinical validation · predetermined change control plans · post-market surveillance · digital pathology · uk medical device regulation

Original article: <https://intuitionlabs.ai/articles/mhra-ai-airlock-phase-2-evidence-case-studies>



In this article

1. Executive Summary
 2. Introduction and Background
 3. Key Changes
 4. Seven Case Studies: Product-to-Evidence Blueprint
 5. Implementation Considerations and Process Changes
 6. AI-Powered IVD Performance Evaluation
 7. PCCPs and Post-Market Evidence
 8. Data Analysis and Evidence
 9. Implications and Future Directions
 10. Conclusion
-

Executive Summary

MHRA AI Airlock Phase 2, the **Medicines and Healthcare products Regulatory Agency (MHRA)** sandbox programme, examined **seven artificial intelligence (AI) technologies across three regulatory challenges**: intended-purpose scope and validation, performance evaluation for AI-powered in vitro diagnostics (**IVDs**), and predetermined change control plans (**PCCPs**) with **post-market surveillance (PMS)**. (www.gov.uk) Selected from **51 applicants**, the cases form a specialist learning portfolio, not a representative survey of AI medical-device performance. (medregs.blog.gov.uk) The programme publication was updated on **27 July 2026** and explicitly states that the report is **not formal MHRA guidance**. (www.gov.uk)

The useful output for developers is a **product-to-evidence blueprint**. Ambient documentation and conversational support raise scope and output-control questions; digital pathology raises comparator, specimen and laboratory questions; summarisation, rare-disease assessment and skin-image applications raise update and monitoring questions. Evidence strength must be labelled accurately. The candidate reports are innovator-authored perspectives, and a multi-environment allocation does not establish live clinical validation. (assets.publishing.service.gov.uk) Proposed validation packages should identify the claim, reference data, clinical metric, human review arrangement and release decision.

Later formal guidance changes how historical scope discussions should be interpreted. MHRA's **29 July 2026 ambient voice technology (AVT) guidance** says unintended outputs and misuse do **not inherently create new intended purposes**. (www.gov.uk) Scope control and qualification assessment should therefore remain distinct. Human oversight needs evaluation over time, rather than being assumed permanently effective. (medregs.blog.gov.uk) International good machine learning practice also evaluates the human-AI team within its intended environment. (www.imdrf.org) For pathology and other diagnostic systems, clinically meaningful evidence extends beyond an aggregate accuracy figure.

The policy position is also time-sensitive. At the report date, **PCCPs were not formally embedded in the United Kingdom (UK) framework**. (assets.publishing.service.gov.uk) On **6 October 2026**, the government accepted **44 National Commission into the Regulation of AI in Healthcare recommendations**, with draft PCCP guidance planned for **December 2026**; acceptance is not an enacted change-control route. (www.gov.uk) For a team entering Great Britain, the immediate recommendation is to build bounded change

protocols, subgroup and site monitoring, and traceability from risk to metric to action, while confirming current assessment expectations. The evidence package should make clear what has been measured, what remains uncertain and what decision each result supports.

Introduction and Background

MHRA AI Airlock Phase 2 is best read as a collection of evidence-design questions. Its publication was issued on **9 June 2026** and updated on **27 July 2026**, including a revision to the engagement route in recommendation 20. It explicitly disclaims formal guidance status. (www.gov.uk)

This report examines what **artificial intelligence as a medical device (AIaMD)** developers entering **Great Britain** can take from the programme: how to specify intended purpose, evaluate clinical performance, document changes and maintain evidence after deployment. The official cohort divided its work among intended-purpose scope and validation, AI-powered in vitro diagnostics (**IVDs**), and post-market surveillance (**PMS**) with predetermined change control plans (**PCCPs**). (www.gov.uk)

The sources have different authority. MHRA's programme synthesis describes sandbox findings; the innovator reports describe candidate-led testing and perspectives. Neither should be presented as a certificate of product approval or an independently replicated comparison of the participants. (assets.publishing.service.gov.uk) (www.gov.uk)

There is also a timeline discrepancy. The publication describes **April 2025 to May 2026**; the programme executive summary describes **April 2025 to March 2026**. These labels should be retained transparently rather than silently reconciled. (www.gov.uk) (assets.publishing.service.gov.uk)

The practical audience spans founders, regulatory and clinical leaders, machine-learning (**ML**) teams, Approved Body readiness teams and National Health Service (**NHS**) deployment partners. The organising question is whether each proposed claim has a defined population, evidence source, acceptance rule and response when the evidence changes.

The implementation perspective is complementary to regulatory review. IntuitionLabs describes its information-layer work in terms of authoritative sources, permissions, retrieval and evaluation; here, that perspective informs evidence traceability rather than a competing medical-device offering. (intuitionlabs.ai) The proposed matrices and checklists below are this report's analytical recommendations. They should be adapted with the responsible clinical, statistical and regulatory professionals, rather than treated as an additional MHRA requirement.

Key Changes

Intended-purpose interpretation now has a later formal source

The sandbox should be read alongside MHRA's formal **ambient voice technology (AVT)** qualification and classification guidance, published on **29 July 2026**. Its administrative examples include reviewed transcription, summarisation and draft discharge documentation that do not derive new clinical recommendations. (www.gov.uk)

Crucially, the guidance says misuse and unintended outputs do **not inherently introduce a new intended purpose**. An unsupported generated statement therefore calls for risk control and scope review, not an automatic conclusion that the device has changed classification. Manufacturers should reassess the application of device regulations with each product change. (www.gov.uk) (www.gov.uk)

This matters when interpreting historical discussions of documentation becoming decision support. The evidence question is what the manufacturer intends the system to do, what its functions and claims support, and how the system is kept within that boundary. A label alone cannot answer the product team's design questions.

Oversight becomes an evaluated lifecycle control

MHRA's official interpretation says human-in-the-loop interactions cannot be assumed static over the device lifecycle. (medregs.blog.gov.uk) The International Medical Device Regulators Forum (**IMDRF**) likewise places human-AI team performance within the intended use environment. (www.imdrf.org)

The analytical consequence is to specify oversight as a workflow: who reviews, what information they see, what they can correct, and what happens when review is unavailable. The review step needs its own evidence, not merely a checkbox stating that a clinician remains responsible.

Performance and change thresholds need clinical meaning

MHRA distinguishes statistical significance from clinical importance. (medregs.blog.gov.uk) The shared Food and Drug Administration (**FDA**), Health Canada and MHRA PCCP principles call for scientifically and clinically justified performance methods and metrics. (www.fda.gov)

Accordingly, a monitoring plan should distinguish detection thresholds, clinical acceptance limits and release decisions. A detectable movement in an aggregate metric does not itself specify whether to retrain, investigate an unnamed process issue, change instructions or pause an update. Those actions require a prespecified rationale.

Existing obligations and proposed reforms remain separate

Strengthened Great Britain PMS requirements came into force on **16 June 2025**. (www.gov.uk) At the Phase 2 report date, PCCPs were not formally embedded in the UK regulatory framework. (assets.publishing.service.gov.uk) The later Commission still described further policy development as necessary to operationalise the approach. (www.gov.uk)

The implication is practical: develop a bounded change plan now, while confirming the applicable assessment and notification route. A useful engineering plan and a formally available regulatory mechanism are different deliverables.

Seven Case Studies: Product-to-Evidence Blueprint

The official cohort identified **three multi-environment candidates** and **four simulation candidates**. Candidate type describes programme allocation, not proof that every allocated environment produced **live clinical evidence**. (www.gov.uk) The candidate-authored reports clarify the testing actually performed.

(assets.publishing.service.gov.uk)

Table 1 combines that allocation with concise product descriptions. Its evidence questions are proposed design priorities, not MHRA acceptance criteria.

Candidate and product	Purpose and challenge	Testing described by candidate	Proposed evidence question
TORTUS	Ambient clinical documentation; intended-purpose scope and validation. The developer distinguishes clinician assessment during development from production monitoring. (tortus.ai)	Multi-environment allocation; shadow-run research using real-world transcripts, without changing normal use. (assets.publishing.service.gov.uk)	Can supported transformations be distinguished from new clinical interpretation, across specialties and configurations?
Numan, Nu & Aegis	Weight-management wellbeing conversation and monitoring; intended-purpose scope. Numan documents a boundary excluding dosing advice. (www.numan.com)	Simulation-only qualitative scope exploration. (assets.publishing.service.gov.uk)	Which input, memory and personalisation changes alter the function being claimed?
Panakeia, PANProfiler Colorectal (MSI/MMR)	Pathology-based colorectal biomarker assessment; AI-IVD performance. The product description concerns microsatellite instability (MSI) and mismatch-repair (MMR) status. (www.panakeia.ai)	Multi-environment allocation; virtual testing and simulation, with no real-world clinical or operational testing reported. (assets.publishing.service.gov.uk)	What reference standard, specimen distribution and laboratory variation support the biomarker claim?
Octopath	Digital pathology cell counting, classification and scoring; AI-IVD performance. Its current website labels the offering research use only. (octopath.ai)	Simulation-only discussion of evidence and claim expansion. (assets.publishing.service.gov.uk)	What evidence transfers between tissues, and what requires tissue-specific evaluation?
NHS England FDP, Safe Summarisation	Summaries of hospital stays, including potential discharge-summary support; PMS and PCCPs. (www.gov.uk)	Multi-environment allocation; synthetic notes in a virtual Federated Data Platform (FDP) environment plus simulation. (assets.publishing.service.gov.uk)	Can content quality, review behaviour and version changes be monitored together?
Eye2Gene	Ranked likely genes supporting inherited-retinal-disease assessment; PMS and PCCPs. (eye2gene.com)	Simulation-only exploration of update and significant-change decisions. (assets.publishing.service.gov.uk)	How should new gene knowledge, imaging variation and delayed outcomes affect update eligibility?
DeepX AI	Skin-lesion image assessment and triage; PMS and PCCPs. Its website describes electronic health record (EHR) integration. (www.deepxhealth.com)	Simulation-only conceptual scenarios, without live deployment. (assets.publishing.service.gov.uk)	How should subgroup evidence and oversight arrangements govern a proposed change?

The matrix supports a planning comparison, not a ranking. Workshop discussion can refine a protocol; virtual testing can characterize behaviour under controlled inputs; shadow-mode work can examine contextual fit. These evidence types should occupy different lines in the evidence register. NICE explicitly recognises silent-mode evaluation as a way to observe local inputs before outputs enter care decisions. (www.nice.org.uk)

MHRA describes TORTUS as providing automated clinical documentation integrated within electronic health records. (www.gov.uk) Octopath's research-use-only statement should remain visible when discussing potential future medical claims. (octopath.ai) UCL's launch account also describes a research interface for digitised tissue specimens. (www.ucl.ac.uk)

The value of the cases is their diversity of evidence questions. None supplies a universal benchmark for all documentation tools, pathology models or retinal applications. The next step is to convert the relevant question into a product-specific protocol with a defined claim and auditable result.

Implementation Considerations and Process Changes

Build a scope decision tree before a validation dataset

The current AVT guidance makes manufacturer-intended function central to qualification. (www.gov.uk) For development teams, a proposed decision tree can begin with the following questions:

- **Output function:** Does the feature reproduce supplied information, structure it, or introduce a new clinical inference?
- **Medical claim:** What diagnosis, treatment, monitoring or other medical benefit is explicitly claimed?
- **User and recipient:** Is the output for clinician review, patient communication or an automated downstream process?
- **Input boundary:** Does the feature introduce laboratory results, longitudinal memory, another data source or a new population?
- **Change consequence:** Does the proposal alter intended purpose, workflow responsibility or the evidence supporting the claim?

The tree should end in a recorded decision: unchanged claim with justified verification, a new claim requiring an evidence plan, or a qualification question for regulatory engagement. It should not end in a generic assertion that all generative systems share the same classification.

Numan's simulation explored **four staged scopes**, illustrating how personalisation and clinical data can be discussed as separate product changes rather than bundled into an undefined conversational capability. (assets.publishing.service.gov.uk) Its current first-party description also documents an explicit dosing-advice boundary. (www.numan.com) These are useful inputs to a scope specification, not proof that a new personalised feature is clinically validated.

Test guardrails against the intended function

Large language models (LLMs) require evaluation methods matched to the output being assured. The programme report rejects LLM-as-a-judge evaluation as the sole assurance mechanism.

(assets.publishing.service.gov.uk) NICE requests identifiable training and validation datasets, including source and version information. (www.nice.org.uk)

A proposed guardrail protocol should specify:

- **Boundary prompts:** Inputs that request functions outside the stated scope.

- **Source fidelity:** Whether clinically material statements remain supported by supplied information.
- **Omission handling:** Whether required information is preserved or an uncertainty is surfaced.
- **Configuration coverage:** Prompt, model, retrieval and interface versions evaluated together.
- **Independent adjudication:** A clinician-defined scoring rubric and a process for resolving disagreement.

These tests should generate reusable artefacts: input provenance, expected handling, generated output, adjudicator decision and the release rule. Preserve disagreements rather than forcing every example into a single binary label. The record should show whether a result is a model assessment, a guardrail assessment or a complete workflow assessment.

IMDRF's good machine learning practice (**GMLP**) principles require appropriate independence between training and test datasets. (www.imdrf.org) An evidence team should therefore make patient, site and acquisition dependencies visible when allocating data, rather than relying on a random row split as its only separation argument.

Evaluate the reviewer as part of the system

The **Developmental and Exploratory Clinical Investigations of Decision-support systems driven by Artificial Intelligence (DECIDE-AI)** reporting guideline asks evaluators to consider patient and user variability. (www.nature.com) The **Standard Protocol Items: Recommendations for Interventional Trials, Artificial Intelligence (SPIRIT-AI)** protocol-reporting extension calls for instructions, required skills and human-AI interaction to be described. (pubmed.ncbi.nlm.nih.gov) (pubmed.ncbi.nlm.nih.gov)

A proposed oversight protocol should record:

- **Reviewer role:** Training, permissions and expected responsibility.
- **Review evidence:** What source information is accessible alongside the generated output.
- **Review behaviour:** Corrections, escalations and task-specific review patterns.
- **Workload context:** Conditions under which the review process is evaluated.
- **Fallback:** The approved workflow when output or review is unavailable.

NICE identifies professional overrides as a monitoring signal. (www.nice.org.uk) Interpretation still requires context: a low override rate could represent appropriate agreement, a workflow change or insufficient observation. Establish the clinical meaning through adjudicated samples and user evaluation.

Research and adoption governance must also be planned. HRA notes that relevant approval and patient-data requirements can apply whether the technology is a medical device or not. (www.hra.nhs.uk) NHS **Digital Technology Assessment Criteria (DTAC)** operate alongside other approvals, while **DCB0129** and **DCB0160**, clinical risk-management standards for digital systems, do not replace medical-device requirements. (digital.nhs.uk) (digital.nhs.uk)

AI-Powered IVD Performance Evaluation

The pathology cases direct attention to the evidence underlying a diagnostic output. IMDRF's software-as-a-medical-device (**SaMD**) clinical-evaluation framework distinguishes **clinical association**, **analytical validation** and **clinical validation**, and treats evaluation as continuous. (www.imdrf.org) These components provide a

useful organising vocabulary rather than a substitute for the applicable IVD framework.

Ground truth means the reference information against which the output is judged. It needs a justification that matches the claim. IMDRF's GMLP principles require fit-for-purpose reference standards. (www.imdrf.org) The Phase 2 report notes that agreement metrics can be more appropriate when the comparator is imperfect. (assets.publishing.service.gov.uk)

For a biomarker inference, the proposed evidence package should distinguish what the software predicts from how the reference status was established. For a cell-scoring application, specify the scoring convention, annotator process and handling of ambiguous material. For a proposed additional tissue claim, explain what changes in the biological question, specimen preparation and clinical use.

A practical performance protocol should address:

- **Reference provenance:** Laboratory method, interpretation process and adjudication of discordance.
- **Specimen coverage:** Relevant tissue, preparation, image quality and exclusion categories.
- **Site independence:** Which laboratories contribute development and evaluation data.
- **Threshold justification:** Why a selected operating point is acceptable for the intended pathway.
- **Clinical utility:** What downstream decision the output is intended to support.

These are proposed protocol fields. They make the claim assessable and expose where further evidence is needed. They also prevent a familiar model metric from substituting for a clinical question.

NICE states that test accuracy alone does not demonstrate clinical utility. (www.nice.org.uk) A laboratory-facing report should therefore describe the role of the result: replacing an existing measurement, triaging samples for another test, or contributing to a clinician's assessment. The evidence needed for those roles may differ even when the underlying algorithm remains unchanged.

Representativeness must be planned across the deployment chain. GMLP includes patient characteristics, settings and measurement inputs within training, testing and monitoring data. (www.imdrf.org) The programme report separately cautions that national benchmarking may not replace local validation. (assets.publishing.service.gov.uk)

The analytical recommendation is to separate **technical transfer** from **clinical transfer**. Successful processing of images from another scanner answers an acquisition question. Evidence supporting a different specimen or decision pathway answers a clinical question. A combined protocol should state which claim each dataset actually supports.

This distinction is especially useful for modular pathology systems. A reusable architecture may support efficient evidence collection, but the technical possibility of reuse does not itself demonstrate clinical suitability for every additional claim. Maintain a claim-by-claim evidence register instead of an undifferentiated statement that the platform has been validated.

PCCPs and Post-Market Evidence

Design the plan without overstating the current UK route

A PCCP describes bounded planned changes and how they will be assessed. The shared regulator principles identify **five** organising principles: focused and bounded, risk-based, evidence-based, transparent, and a total-product-lifecycle perspective. (www.fda.gov)

The FDA's final **August 2025** guidance describes planned modifications, the methodology for developing, validating and implementing them, and impact assessment. Under its US mechanism, **modifications within an authorised plan** need not each trigger another marketing submission. (www.fda.gov) (www.fda.gov) That US consequence should not be attributed automatically to Great Britain.

IMDRF's **August 2026** PCCP document provides international content principles, while expressly stating that it does not replace jurisdictional laws. (www.imdrf.org) (www.imdrf.org) For a UK submission, establish the current legal route and the assessing body's expectations before describing a change as pre-authorised.

Table 2 translates the bounded-plan concept into proposed evidence controls. It is an engineering and regulatory-discussion template.

Proposed modification	Protocol and boundary	Acceptance evidence	Action when boundary is not met
Retraining for the existing claim	Define eligible data, pipeline and fixed purpose. Shared principles bound changes by intended use. (www.fda.gov)	Claim-specific performance and justified subgroup criteria.	Hold release; document the assessment route for the proposed change.
Prompt or retrieval update	Test the complete configuration against the output specification.	Source fidelity, required-content coverage and adjudicated workflow results.	Restore the supported configuration; revise the protocol.
Acquisition or site extension	Document input compatibility and deployment context.	Independent site evidence and fit-for-purpose reference assessment. (www.imdrf.org)	Restrict the supported configuration pending additional evidence.
Oversight modification	Treat reviewer role and autonomy as explicit design variables.	Human-AI team evaluation in the intended environment. (www.imdrf.org)	Retain the established review arrangement.
Proposed claim expansion	Record new population, tissue, clinical role or purpose.	Evidence specifically supporting the additional claim.	Seek the applicable qualification and conformity assessment decision.

The table separates a change description from permission to implement it. Shared PCCP principles envisage mechanisms to detect and stop or revert changes that miss prespecified criteria. (www.fda.gov) The September Commission similarly warns that significance cannot be judged from the technical modification alone. (www.gov.uk)

For internal review, ask whether the proposal changes purpose, population, setting, autonomy, oversight, benefit-risk assumptions or cumulative behaviour. Record the reasoning even when the model architecture is unchanged. A configuration diff is useful evidence, but it is not the complete clinical assessment.

Build monitoring around decisions

The programme report recommends linking monitored thresholds to whether changes proceed, pause or terminate. ([assets.publishing.service.gov.uk](#)) WHO's regulatory-considerations overview also addresses performance evaluation and monitoring, while expressly presenting itself as an overview rather than policy or a regulatory framework. ([www.who.int](#)) ([www.who.int](#))

For operational planning, define a monitoring matrix across:

- **Population:** Clinically relevant demographic and disease subgroups.
- **Site:** Laboratory, trust, acquisition platform and deployment configuration.
- **Behaviour:** Output characteristics, clinician corrections and workflow use.
- **Outcome:** Available reference labels, clinical results and label-delay patterns.
- **Version:** Model, prompt, retrieval corpus and associated release documentation.

These dimensions are proposed controls, not a requirement to collect every possible variable. Select them from the intended purpose and risk analysis, with a clear purpose for each item.

NICE calls for processes to detect declining performance within groups over time. ([www.nice.org.uk](#)) For skin-image systems, a proposed subgroup plan should explicitly examine relevant skin-tone representation and acquisition conditions. For rare-disease applications, it should address sparse outcomes and evolving reference knowledge. These are design recommendations, not reported subgroup-effect estimates from the sandbox.

Eye2Gene's website describes activity logging and role-based access, which can support an auditable feedback process. ([eye2gene.com](#)) TORTUS describes production observation separately from development-time clinician assessment. ([tortus.ai](#)) Both are first-party capability statements; neither establishes that every deployment has the same monitoring coverage.

Data Analysis and Evidence

What the available denominators support

Phase 2 selected **seven technologies from 51 applicants**. ([medregs.blog.gov.uk](#)) The simple ratio, $7 \div 51 = 13.7\%$, describes cohort selection, not device success, clinical effectiveness or the prevalence of regulatory challenges. ([medregs.blog.gov.uk](#)) Application portfolios are purposive evidence sources.

The candidate-authored Panakeia case study describes **791 patients** from St James's University Hospital, Leeds. ([assets.publishing.service.gov.uk](#)) It reports **97.86% sensitivity**, **93.82% specificity** and **99.65% negative predictive value (NPV)**. ([assets.publishing.service.gov.uk](#)) These figures belong to that evaluation, not a pooled Phase 2 benchmark. Their interpretation requires the comparator, disease distribution, exclusions and intended clinical role; the headline percentages should not be transferred to another laboratory or use case.

A separate published Eye2Gene study used an external test dataset of **39,596 retinal scans from 836 patients**, with external validation across **five clinical centres**. ([www.nature.com](#)) ([www.nature.com](#)) This is published research evidence, not the sandbox's simulation result. The distinction between image count and

patient count is material when planning independence and precision.

Reporting tools also have defined scopes. DECIDE-AI includes **17 AI-specific** and **ten generic** reporting items for early-stage evaluation, while **Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, Artificial Intelligence (TRIPOD+AI)** provides a **27-item** reporting checklist.

(www.nature.com) (www.tripod-statement.org) **Consolidated Standards of Reporting Trials, Artificial Intelligence (CONSORT-AI)** concerns reporting trials evaluating interventions with an AI component.

(www.equator-network.org) Checklist completion makes evidence easier to inspect; it does not establish regulatory conformity.

Sample-size planning requires project inputs

There is no monitoring sample-size recommendation in this report. A statistical plan should begin with user-supplied inputs:

- **Endpoint:** The particular binary proportion or other outcome being monitored.
- **Expected rate:** A justified planning value, not an assumed universal performance level.
- **Precision or detectable change:** The uncertainty or movement the study must address.
- **Confidence and power:** Prespecified statistical error limits where relevant.
- **Strata and dependencies:** Site, subgroup, patient and repeated-measure structure.

For approximate precision planning of an independent binomial proportion, the familiar expression is $n \approx z^2 p(1 - p) / e^2$, where the user supplies the expected proportion p , half-width e and confidence quantile z . The **National Institute of Standards and Technology (NIST)** presents this normal-approximation approach with explicit assumptions. (itl.nist.gov) It should be calculated separately where a stratified design requires stratum-specific precision. (itl.nist.gov)

Precision estimation differs from powered detection of a prespecified change. NIST's change-detection calculation takes the baseline and target proportions, significance and power into account. (www.itl.nist.gov) Rare outcomes, small samples and proportions near the boundary need particular care: NIST cautions that symmetric normal-approximation intervals can be inaccurate with small event counts or sample sizes. (www.itl.nist.gov)

The proposed planning rule is to commission the statistical design before treating a monitoring window as adequate. Record the unit of analysis, missing-label strategy, clustering, repeated looks and subgroup uncertainty. Otherwise, a dashboard may display stable percentages without demonstrating that the planned clinical change could be detected.

Implications and Future Directions

Trace risks through to accountable action

Table 3 proposes a traceability structure connecting the intended-purpose statement to evidence, alerts and **corrective and preventive action (CAPA)**. It is a design template, not a record of any participant's operational results.

Risk or evidence question	Metric and evaluation unit	Proposed alert rule	Action and retained record
Unsupported clinical content	Clinician-adjudicated source fidelity by reviewed encounter.	Prespecified clinical acceptance limit, with uncertainty reported.	Review affected configuration; retain source, output and adjudication.
Subgroup uncertainty	Claim-specific performance and reference coverage by relevant subgroup.	Insufficient evidence or breached subgroup criterion.	Expand evidence collection or restrict supported use; retain rationale.
Site transfer	Performance by acquisition and deployment configuration.	Local acceptance criterion not met.	Hold the extension; retain site verification and decision.
Oversight adequacy	Review and correction behaviour linked to adjudicated samples.	Clinically justified change in oversight effectiveness.	Reassess interface, training or review arrangements.
Planned update eligibility	Prespecified performance and scope checks by release version.	Any release condition outside the bounded plan.	Stop the update; document change assessment and any restoration.

The interpretive point is ownership. Every alert needs a decision-maker, timeframe and closure record. Clinical evaluation is an iterative part of quality management, according to IMDRF. (www.imdrf.org) DTAC remains an additional deployment assessment rather than a replacement for device approvals. (digital.nhs.uk)

Engage with a concrete evidence package

For a near-boundary product, the revised programme recommendation suggests engaging MHRA through info@mhra.gov.uk before deployment. (assets.publishing.service.gov.uk) The package should include representative outputs and a user journey, not only a technology description.

The Commission describes the current Great Britain approach for medium and high-risk devices as conformity assessment by MHRA-designated Approved Bodies, rather than direct conformity assessment by MHRA itself. (www.gov.uk)

For an assessing-body discussion, distinguish the study design from its reporting format. DECIDE-AI does not prescribe a fixed study design. (www.nature.com) SPIRIT-AI asks for the instructions and skills needed to use the intervention. (pubmed.ncbi.nlm.nih.gov) Performance criteria should have scientific and clinical justification, (www.fda.gov) while development and evaluation data need appropriate independence. (www.imdrf.org)

Proposed engagement questions are:

- **Qualification:** Which example outputs and user journeys resolve the boundary question?
- **Validation:** Which comparator, sites and populations must the evidence cover?
- **Change:** Which proposed modifications require a new assessment or notification?
- **Oversight:** What demonstrates that the reviewer arrangement works in context?
- **Surveillance:** Which metrics and subgroup evidence support continued conformity?

A proposed pre-submission checklist is:

- **Purpose:** Intended users, population, clinical role and excluded functions.

- **Claims:** A mapping from each claim to its supporting evidence.
- **Data:** Provenance, reference standards and separation of development from evaluation.
- **Oversight:** Reviewer instructions, workflow evidence and fallback.
- **Change:** Proposed modifications, acceptance criteria and regulatory decision route.
- **Monitoring:** Subgroups, sites, labels, triggers, actions and accountable owners.
- **Deployment:** Research governance, local assurance and budget impact.

HRA distinguishes research from service evaluation under its policy definitions, so the actual activity should determine the approval question. (www.hra.nhs.uk) NICE's user guide also asks for budget-impact analysis; clinical evidence and adoption economics should therefore be planned together. (www.nice.org.uk)

As of **13 October 2026**, Phase 3 applications are open following the **6 October** update. (www.gov.uk) The government accepted **44 Commission recommendations** and committed to draft PCCP guidance for public comment by **December 2026**, with an implementation roadmap by **Spring 2027**. (www.gov.uk) The promised change framework is intended to define allowable-change boundaries, but its future details remain to be consulted on. (www.gov.uk)

For an adjacent implementation advisor, the durable priority is an operable evidence process. IntuitionLabs' own measurement description includes quality, risk signals, reliability and support burden. (intuitionlabs.ai) Those operational measures can complement the product's clinical evidence without replacing the manufacturer's responsibilities or the relevant assessment.

Conclusion

The practical answer to **MHRA AI Airlock Phase 2 evidence requirements** is a product-to-evidence plan, with careful separation between formal requirements, sandbox observations and proposed implementation controls. The programme report remains explicitly non-guidance. (www.gov.uk) Its cases help teams frame questions; they do not establish a shared regulatory outcome for every product using a similar model.

Start with the intended function and clinical role. Identify what the output changes in the workflow, which population and setting the claim covers, and how that claim will be assessed. Then specify the reference standard, independent evaluation data, reviewer arrangement and evidence needed for local adoption. Preserve the distinction between a discussion workshop, a controlled evaluation and evidence from actual use.

For evolving systems, connect the change proposal to the existing claim and to release conditions. A PCCP-inspired plan can make development more disciplined, while the applicable jurisdictional route still needs confirmation. International principles expressly retain that jurisdictional distinction. (www.imdrf.org)

For post-market evidence, prefer an accountable decision process over a collection of dashboard metrics. Define the relevant subgroup and site questions, label availability, statistical uncertainty and actions that follow a breached condition. Evaluate whether human oversight continues to serve its intended function.

The recommended readiness test is whether an assessor or NHS partner can trace a claim from purpose to data, metric, threshold, action and retained decision. That trace is the useful deliverable: evidence that can be inspected, updated and acted on as the product and its deployment context evolve.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. <https://intuitionlabs.ai/articles/post-market-surveillance-ai-locked-continuous-learning>
2. <https://www.gov.uk/government/publications/ai-airlock-phase-2-cohort>
3. <https://medregs.blog.gov.uk/2026/06/09/advancing-ai-regulation-in-healthcare-insights-from-ai-airlock-phase-2/>
4. <https://www.gov.uk/government/publications/ai-airlock-sandbox-phase-2-programme-report>
5. https://assets.publishing.service.gov.uk/media/6a59e205908b618a53702f8b/Phase_2_Candidate_Case_Studies_for_publication.pdf
6. <https://www.gov.uk/government/publications/ambient-voice-technology-enabled-products/ambient-voice-technology-enabled-products>
7. https://www.imdrf.org/sites/default/files/2025-02/IMDRF_AIML%20WG_GMLP_N88%20Final.pdf
8. https://assets.publishing.service.gov.uk/media/6a676af523e57bd0a0ea6852/AI_Airlock_Phase_2_Programme_Report_v2.pdf
9. <https://www.gov.uk/government/publications/government-response-to-the-national-commissions-recommendations-on-the-regulation-of-ai-in-healthcare/government-response-to-the-national-commissions-recommendations-on-the-regulation-of-ai-in-healthcare>
10. <https://intuitionlabs.ai/articles/samd-definition-and-scope>
11. <https://intuitionlabs.ai/>
12. <https://www.fda.gov/medical-devices/artificial-intelligence-enabled-medical-devices/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles>
13. <https://www.gov.uk/government/publications/implementation-of-the-future-regulation-of-medical-devices/implementation-of-the-future-regulations>
14. <https://www.gov.uk/government/publications/national-commission-into-the-regulation-of-ai-in-healthcare-recommendations-for-a-future-regulatory-framework/national-commission-into-the-regulation-of-ai-in-healthcare-recommendations-for-a-future-regulatory-framework>
15. <https://intuitionlabs.ai/articles/ai-samd-clinical-evidence-real-world-studies>
16. <https://tortus.ai/blog/running-avt-as-a-class-ii-a-device>
17. <https://www.numan.com/inside-numan/safety-foundation-ai-healthcare>
18. <https://www.panakeia.ai/post/panakeia-launches-panprofiler-colon-an-ai-driven-tool-for-precision-medicine-in-colon-cancer>
19. <https://octopath.ai/>
20. <https://eye2gene.com/>
21. <https://www.deepxhealth.com/>
22. <https://www.nice.org.uk/corporate/ecd7/chapter/how-to-meet-the-standards>
23. <https://www.ucl.ac.uk/engineering/news/2024/dec/new-ai-cancer-pathology-research-platform-launched>
24. <https://www.nature.com/articles/s41591-022-01772-9>
25. <https://pubmed.ncbi.nlm.nih.gov/33328049/>
26. <https://www.hra.nhs.uk/planning-and-improving-research/research-planning/how-were-supporting-data-driven-technology/overview-legal-requirements-using-health-and-care-data-development-and-deployment-data-driven-technologies/4-checklist-standards-codes-and-approvals/>

27. <https://digital.nhs.uk/services/digital-technology-assessment-criteria-dtac>
28. <https://digital.nhs.uk/services/clinical-safety/applicability-of-dcb-0129-and-dcb-0160/nhs-digital-recommendation>
29. <https://intuitionlabs.ai/articles/clinical-evidence-requirements-ai-diagnostics>
30. https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-170921-samd-n41-clinical-evaluation_1.pdf
31. <https://intuitionlabs.ai/articles/fda-predetermined-change-control-plan-pccp-guide>
32. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence>
33. <https://www.imdrf.org/sites/default/files/2026-08/IMDRF%20SaMD%20WG%20N90%20FINAL%20Essential%20Principles%20and%20Content%20PCCPs%20%281%29.pdf>
34. <https://www.who.int/westernpacific/publications/i/item/9789240078871>
35. <https://www.nature.com/articles/s42256-025-01040-8>
36. <https://www.tripod-statement.org/scope/>
37. <https://www.equator-network.org/reporting-guidelines/consort-artificial-intelligence/>
38. <https://itl.nist.gov/div898/handbook/ppc/section3/ppc333.htm>
39. <https://www.itl.nist.gov/div898/handbook/prc/section2/prc242.htm>
40. <https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/exacbi.htm>
41. <https://www.nice.org.uk/corporate/ecd7/resources/evidence-standards-framework-for-digital-health-technologies-user-guide-9024813901/chapter/how-to-use-the-framework>
42. <https://www.gov.uk/government/collections/ai-airlock-a-regulatory-sandbox-for-ai-enabled-medical-devices>

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore [AI enablement](#), [pharma and biotech software development](#), [data engineering and BI](#), and [Veeva services](#). [Contact IntuitionLabs](#) to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: <https://intuitionlabs.ai>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.