Medicines & Healthcare products Regulatory Agency

PHG FOUNDATION
making science work for health

Insights from the AI Airlock Simulation Workshop: Post Market Surveillance and Predetermined Change Control Plans

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

Contents

Overview

Introduction

This report covers key findings from a workshop on *post market surveillance (PMS) and predetermined change control plans (PCCPs)* held on 11th February 2026. This simulation workshop – an interactive, focussed roundtable workshop 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. Medical device regulation applies across the full product lifecycle, from pre-market evaluation through to post market monitoring.

For SaMD and AIaMD, lifecycle management can be more complex due to the likelihood of the device needing or receiving relatively frequent post market updates. Many modern AI systems are probabilistic, rather than deterministic and their outputs are shaped by data, deployment context and model behaviour in ways which may be difficult to fully anticipate or trace. Such updates could include model retraining, data updates, performance optimisation, or changes to system integration which may have downstream effects on clinical outputs that are not immediately apparent as the product's intended purpose may remain unchanged. Therefore, it is particularly important for manufacturers to define, in advance, how changes will be managed and controlled. Current regulatory approaches to change management are based, in part, on determining whether modifications are significant and require regulatory reassessment. For AI systems, this creates challenges at both a regulatory and technical level. The regulatory concept of a "significant change" is not always simple to define for adaptive systems where the boundary between optimisation and substantial alteration may not be so clear. Technically, the probabilistic nature of AI outputs could mean the effect of a change may only become apparent in actual deployment over time, making it difficult to assess impact and significance ahead of real-world use.

PCCPs allow manufacturers to set out, ahead of deployment, the types of changes to a device that may be made, the limits within which they can occur, and how they will be validated. This allows certain updates to be implemented without full reassessment, provided they remain within an agreed scope. While PCCPs are not yet formalised in UK MDR 2002, the MHRA supports their use and aligns with international principles.

PMS provides the ongoing regulatory mechanism for monitoring real world safety and performance. It includes incident reporting, performance monitoring, user feedback, post market clinical follow-up, and corrective and preventive actions. For AIaMD, PMS is particularly important for identifying issues not fully evident at pre-market stage, such as performance drift, data shift, or subgroup disparities. Recent updates to UK MDR 2002 have strengthened PMS requirements, including setting more formal PMS planning, enhanced reporting obligations, and clearer expectations for risk management and corrective action.

Existing processes are largely designed for technologies with fixed functionality, whereas AI systems are often designed to iterate and modify performance over time. Updates such as retraining, refinement, and parameter changes may also be necessary to maintain safety and performance, particularly as real-world conditions evolve. While the current regulatory framework provides a strong foundation for lifecycle, risk, and post market management, practical challenges arise when applying these management functions to frequently changing AI medical devices. Together, PCCPs and PMS are intended to address some of these challenges and support controlled iteration while maintaining assurance of safety and performance over time. This creates several regulatory questions about device updates: which updates can be managed within existing regulatory mechanisms, where could PCCPs be implemented to manage changes, and which constitute a significant change requiring reassessment? This challenge is particularly acute for large language model (LLM) enabled devices, where ongoing updates may be needed to maintain clinical utility. Unlike deterministic software, the effects of a change to an AI system may not always be predictable or verifiable in advance. Unexpected shifts in outputs or performance may only become apparent in real-world use. Therefore, robust continuous monitoring is needed pre and post change implementation to establish a reliable performance baseline and detect meaningful shifts once changes are made. Under current rules, even limited changes may trigger reassessment to manage these risks, which may generally support safety but could also potentially slow proportionate improvement or even adequately maintained performance. PCCPs offer a potential solution by allowing predefined updates within an agreed scope and validation protocol. However, current guidance from the MHRA remains high level, has not yet been placed in practice, and does not fully address the complexities of adaptive and generative AI systems.

The objectives of this simulation workshop on *PMS and PCCPs* were to:

- Explore how PCCPs can be defined, structured, and operationalised for LLM enabled devices
- Explore the possibility of integrating PMS and PCCPs effectively across the AIaMD lifecycle
- Identify evidence, validation criteria, and assurance requirements for a PCCP
- Explore a suitable definition for “significant change”
- Support the development of future regulatory guidance for PCCPs

The workshop brought together a diverse group of stakeholders and three real-world candidates: NHS Safe Summarisation (a Multi-environment candidate), DeepX AI (a simulation candidate) and Eye2Gene (a simulation candidate). Stakeholder expertise spanned multiple sectors, including healthcare, academia, regulatory and compliance, policy and ethics, research and innovation, AI specialists, and NHS England.

Multi-environment candidate: NHS Safe Summarisation (NHS SS)

The product is an LLM-based clinical document summarisation tool designed to produce structured summaries of patient notes from electronic healthcare records for clinician use. The key regulatory questions were around defining PCCP criteria for AIaMDs, specifically LLM-based summarisation tools, to generate insights into best practices for continuous regulatory compliance of evolving AI medical devices.

Simulation candidate: DeepX AI

DeepX AI's Dermosight is an AI-based teledermatology triage tool that uses specialist imaging hardware to provide clinician-in-the-loop decision support within a pharmacy-based pathway for assessment of skin lesions suspected of skin cancer. The key regulatory questions for the workshop were to define the necessary structure, evidence and regulatory acceptance criteria for PCCPs for AIaMD, and what constitutes significant change.

Simulation candidate: Eye2Gene

The product, currently at the research use stage, is an AI-powered tool which analyses retinal scans in combination with genomics data to support identification of patients with inherited retinal disease. The key regulatory question for Eye2Gene was establishing an effective PMS system for rare disease conditions, where prevalence is low, pathways are long, prospective evidence may be difficult to generate and exploring the ideal template and content of a PCCP to manage changes effectively.

Key insights from the workshop

Defining significant change in AIaMD

The first session discussed and explored what constitutes a "significant change" in AIaMD within the current Great Britain regulatory context, where no formal AI-specific guidance is established. The discussion aimed to build a shared understanding to support future policy development for adaptive AI and lifecycle-based regulation. Participants discussed and explored the role of PCCPs as a mechanism to manage certain pre-specified changes without full regulatory reassessment, while recognising that PCCPs cannot cover changes affecting intended purpose or risk classification. Defining the boundary for which product and environmental changes that can be managed within a PCCP and those that require

reassessment was identified by delegates as a key challenge, making a clear and practical definition of “significant change” essential to the effective use of PCCPs.

Testing a significant change framework

To further understand how significant change could be defined in practice, and to test a structured pathway for identifying such changes, participants then discussed a draft flowchart for identifying significant change developed by MHRA that was intended to potentially support decision-making on what constitutes a significant change for AIaMD. The framework outlined three stages: intended purpose, product and model-level assessment, impact assessment and regulatory consequences. Overall, delegates felt that significance cannot be determined from technical changes alone; it depends heavily on clinical context, user interaction, intended use, and ultimately how changes impact product performance over time.

Stage 1: Intended purpose

The first stage of the flowchart focused on whether a modification changes a device’s intended purpose. Intended purpose was seen as the key regulatory anchor, but a change to intended purpose was not the only factor delegates identified as important to determine a significant change.

There was concern from the delegates that broad intended use wording may not facilitate identifying significant functional changes. Discussion focused on definitions of intended population, intended user, intended environment, structure and function, and how these concepts interact with software evolution, off-label use, and governance frameworks. Participants also explored how intended purpose intersects with other regulatory regimes, including research governance and data protection.

For LLM and adaptive AI enabled devices, defining intended purpose was felt to be more challenging due to the potential of these products to have evolving behaviour and broad use contexts. There was agreement by the delegates that post-market monitoring is essential for such devices, though uncertainty remained on handling adaptive systems and underlying model changes that may occur without clear visibility.

Intended population was seen as a key issue, as using a tool in populations that are underrepresented in training data can alter the risk benefit for the product. This raised questions about what constitutes a meaningful change when the original intended use is broad, such as “all patients,” and whether expanding to new populations under the same validation framework falls within the PCCP scope. Perspectives were conflicting among the delegates however extending use to significantly different populations (e.g. from adult to paediatric use) was agreed as likely a significant change requiring reassessment.

Changes in intended users or deployment context, particularly shifts from professionals to lay users, were widely seen by delegates as significant. Similarly, participants felt that both technical and clinical operating environments can alter risk and should be considered, although the extent to which external changes, such as deployment context, could be included in an intended use statement was unresolved.

Finally, delegates felt that reasonably foreseeable misuse should be accounted for within risk management – a consideration with implications for PMS design where monitoring mechanisms need to be capable of detecting when real-world use diverges from intended purpose. Delegates also noted that intended functionality changes, or real-world divergence from intended purpose may trigger governance frameworks beyond medical devices and some examples highlighted included research ethics and data protection. Overall, delegates indicated that clearer guidance on drafting intended purpose and evaluating significant changes would be helpful, especially for broadly defined claims or incremental expansions which have direct bearing on how PCCP boundaries are defined.

Stage 2: Product/ Model changes

The second stage of the flowchart examined how product and model changes could be assessed, and discussion emphasised that “significance” of a change is context-dependent and should be judged by impact on a particular device’s benefits and risks, user interactions and clinical outcomes.

Introducing new data to support a new population was generally viewed as a significant change, while continuous training updates raised concerns about whether this would effectively trigger constant “significant” changes. In contrast, expanding training data within an already validated population could be considered non-significant, depending on how validation is handled. Regardless, the participants felt that performance changes were important and not to be viewed in isolation; subgroup performance, downstream effects of the change, and an understanding of the device’s overall benefit-risk profile were considered critical context.

Discussion also considered that interface, operating system and usability changes could appear minor but affect user behaviour and safety. Human factors including usability, user experience, and accessibility standards were highlighted as critical considerations, particularly where changes alter how outputs are perceived, require retraining, or impact different user groups; examples such as decimal point shifts or voice-to-text input illustrate potential risks and can introduce performance and usability challenges.

Changes to human oversight were also considered important: reducing oversight or increasing automation was thought more likely to be significant, but how such changes are classified remains unclear.

Finally, effective, ongoing monitoring was seen by delegates as essential for managing change under a PCCP. Participants felt that changes that cannot be adequately monitored should not be eligible for management within a PCCP, though minimum monitoring requirements are not presently defined.

Stage 3: Impact assessment

The third stage of the flowchart explored the impact assessment that should be carried out for a change, emphasising that significance of a change should be linked to a proportionate regulatory response. Rather than simply categorising a change, participants argued that the focus should be on the necessary response: whether that involves a full reassessment, a targeted review, or more proportionate oversight mechanisms. There was agreement from the delegates that changes that would increase risk classification should fall outside the scope of allowable changes in a PCCP and require formal reassessment. There were varying perspectives on the extent of regulatory involvement; some believed most changes should require oversight, whereas others supported a more balanced approach, suggesting that even significant modifications could be handled with lighter regulation.

Participants noted that it can be difficult to predict the impact of changes before deployment. While safety can be assessed through structured risk frameworks such as ISO 14971 identifying concepts of harm, residual risk and benefit-risk balance, performance is harder to evaluate in advance because degradation may only emerge when the device encounters population variation or clinical workflows that differ from controlled conditions. Participants supported aligning assessments with ISO 14971 concepts, including both physical and non-physical harms, though thresholds for triggering reassessment remained unclear. Changes affecting clinical decision-making, interpretability or human oversight were also seen by delegates as potentially significant, even without known performance shifts.

There was strong agreement that cumulative change needs explicit consideration, as while a single minor update may not be significant, multiple incremental changes can collectively alter a device’s risk profile, performance, and clinical impact. Participants highlighted uncertainty around when such accumulation should trigger reassessment and at what point a product effectively becomes “new.” As a result, there was clear support for establishing mechanisms and guidance to monitor and evaluate cumulative impact over time. Some of this kind of assessment and monitoring could potentially be evaluated via PCCPs, as part of an impact assessment. It was also felt that greater clarity is needed on when bug fixes, cybersecurity updates and corrective actions move beyond routine maintenance into significant change.

Finally, participants emphasised the importance of international alignment, particularly in consideration of EU MDR and NHS safety standards, though the extent to which the UK should align or diverge was not discussed in detail.

PCCPs in practice

PCCPs were presented as a mechanism to support planned, bounded and well-characterised changes to AIaMD over time, enabling certain predefined updates without requiring full reassessment. It was emphasised that PCCPs must clearly define the scope of permitted changes, include robust validation and monitoring processes, and maintain regulatory oversight rather than reduce it.

Three real-world case studies were examined during the workshop. Common themes that emerged in the discussion included the importance of clearly defined intended purpose, the difficulty of setting boundaries for acceptable change, the need for effective monitoring and performance thresholds, and the challenge of balancing innovation with patient safety, equity and public confidence.

Managing change in AI-supported patient triage: DeepX AI

Equity and data representativeness

A significant concern raised in this session was the limited availability of representative training data for patients with darker skin tones within UK datasets. In dermatology, where visual assessment is central to the diagnostic task, a model trained predominantly on data from lighter-skinned populations may perform measurably less well for underrepresented groups. This creates an equity risk that is not always visible in aggregate performance metrics.

Discussion highlighted that equity considerations are most effectively addressed at the design stage rather than through post-market remediation, but that it is important that post-market surveillance is capable of detecting subgroup performance disparities as they emerge in real-world use. There was support in principle for PCCPs that include predefined pathways for controlled retraining aimed at addressing identified performance gaps across subpopulations, provided changes remain within pre-specified performance limits and do not alter the device's intended purpose. The use of external or synthetic data to improve representativeness was also discussed as a potentially useful tool, but participants cautioned that synthetic data carries risks of bias amplification and requires careful evaluation.

The boundary between PCCPs and post-market surveillance

A recurring tension in this session concerned the appropriate division of responsibility between PCCPs and PMS. These two mechanisms serve related but distinct functions: PMS identifies what is happening in real-world deployment, whilst a PCCP provides a pre-agreed framework for implementing anticipated changes. Participants noted that for PCCPs to be

operationally useful, there must be clarity about whether predefined performance thresholds can activate a PCCP-governed update without triggering a new regulatory submission. In practice, PMS systems could continuously generate and assess real-world performance data, safety signals, and evidence of clinical use to inform whether planned changes remain justified and safe to implement or can no longer be implemented in accordance with the PCCP and should potentially trigger additional regulatory oversight. This interaction could be achieved by establishing predefined thresholds and monitoring criteria within the PCCP, linking change triggers or conditions (to cease implementing changes or enact responsive changes) to specific PMS outputs such as adverse events, performance drift, emerging risks, or population shifts, noting that PCCPs are not generally intended as a mechanism for responding to device failures. Thus, defining the boundaries of acceptable lifecycle evolution, clarifying when regulatory reassessment is needed, and exploring how real-world performance data should inform ongoing change will help the manufacturers to address the regulatory challenge of managing changes for adaptive or otherwise rapidly modifiable devices. In the absence of such clarity, manufacturers face uncertainty that may inhibit beneficial updates or lead to inconsistent practice.

The participants felt the implications were that:

- Intended use remains central to determining PCCP boundaries. PCCPs can enable controlled evolution but will only do so safely and effectively within clearly defined evidentiary and performance limits. For example, expansion across populations, genes or care settings may trigger regulatory re-evaluation if such changes alter intended use or risk profile and therefore go beyond the bounds of a PCCP.

- There are broader potential challenges in applying PCCPs to AI systems. For AI, retraining can improve performance without changing the intended purpose or architecture, making it unclear whether an update constitutes a change requiring PCCP governance or a post-market clinical follow-up activity. The ambiguity is more pronounced for AI given the relationship between a technical change and the clinical effect may not always be easily traceable in advance.

- A further challenge discussed was the cumulative effect of individually bounded discrete changes. A series of updates that individually fall below the threshold for regulatory significance could, in aggregate, constitute a significant change or materially shift a product's risk profile. Delegates noted that manufacturers should be actively managing and tracking this rather than treating each change in isolation. This is particularly relevant for AI, where adjustments to output framing or workflow integration could gradually alter the degree to which a clinician exercises independent judgement. A well-designed PCCP should explicitly account for cumulative change, defining how sequential updates are tracked, against specific performance metrics, and at what point their

aggregate effect requires reassessment, providing manufacturers with a structured mechanism to govern this proactively.

- The DeepX case study also surfaced a more specific regulatory challenge at the intersection of PCCPs, PMS triggers, data acceptability, and equity. The question arose as to whether a PCCP could pre-agree a pathway for using international or synthetic data to address known performance gaps in underrepresented populations where sufficient UK data is unavailable. Training an AI model to perform reliably across demographic subgroups requires data volumes that may be an order magnitude larger than what is domestically available. This presents a challenge on whether to limit the product's intended population or use international data or synthetic data whose regulatory acceptability is not yet established. DeepX noted that a PCCP that pre-agrees a pathway for evidence-driven remediation as data becomes available would provide a potentially proportionate and workable mechanism. Participants felt there should be risk proportionate regulatory mechanisms available to allow manufacturers to address such disparities when identified after deployment.

Managing change in LLM-based clinical summarisation: NHS Safe Summarisation

This session explored regulatory challenges that are distinctive to generative AI systems and that are not fully addressed by PCCP guidance that has been developed for more conventional AI applications.

Defining significant change for LLM based system

LLM-based systems are architecturally different from traditional AI models. Modifications can occur at multiple layers: the underlying foundation model; the rules or workflow that controls how the model is used, the contextual information it's given and the settings it runs with. Each layer can affect system behaviour, and changes at one layer may have indirect effects on others. This makes it difficult to apply a simple categorisation of "minor" versus "major" software change and to define the relationship of such changes with "significance," since the significance of a modification depends not only on the specific technical changes made, but on the interaction of the change with the rest of the system, and ultimately its impact on the outputs.

Participants agreed that prompt refinements and minor model workflow adjustments are likely to be lower risk and may be manageable within a PCCP, provided they are well-documented and remain within pre-specified bounds of performance. Switching to a different version of a foundation model in a device was viewed as higher risk and delegates felt such changes were likely to fall outside PCCP scope in most cases. The addition of trusted documents or policies into the contextual information provided to the model was seen as a

potentially important quality improvement to the device but one likely still requiring structured evaluation.

Validating probabilistic outputs

Generative AI-based systems are non-deterministic, meaning they can produce outputs that vary in response to the same input. This creates distinctive challenges for both pre-market evidence and ongoing post-market monitoring. Conventional regression testing, which anticipates consistent input-output relationships, is not straightforwardly applicable. Validation must instead account for output variability as a design characteristic, and performance metrics should be defined in ways that are meaningful to the specific product and intended use. As an example for summarisation tasks specifically, including sensitivity to hallucination, omission of clinically relevant information, and inappropriate emphasis.

The session also discussed model-based evaluation, where a separate AI system is used to assess the quality of outputs from the primary model. Participants saw practical value in this approach but noted its limitations, including the risk that the evaluating model may share blind spots with the system being evaluated.

Human oversight and overreliance

The NHS Safe Summarisation tool described, retains human clinicians in the review loop, with the ability to edit summaries, add information, and provide quality ratings. Participants questioned whether this arrangement reliably functions as an effective safeguard against AI errors in practice, particularly over time. As users become more familiar and confident with the system, the thoroughness of review may diminish, potentially reducing the effectiveness of human oversight as a risk mitigation. This risk of overreliance was seen by the delegates as systemic rather than individually attributable, arising from the interaction between the system, the clinical environment, and workflow pressures.

Participants agreed that manufacturers bear responsibility for designing systems in ways that actively support meaningful, ongoing review of outputs rather than merely providing nominal human-in-the-loop arrangements. Monitoring signals that might indicate diminishing engagement, such as reductions in elapsed review time or declining rates of feedback activity, were discussed as potentially useful indicators of behaviour change indicating reduced oversight, though their interpretation requires care given the range of other factors that influence clinician behaviour.

Operational context including interoperability and rollback

Participants emphasised that AI system performance and safety are highly dependent on operational context, including interoperability with EHR systems, infrastructure, and local clinical workflows, and recommended that these factors be explicitly considered within technical documentation. Changes in deployment conditions, data inputs or system

integrations can affect outcomes even without modification to the core product. A PCCP should govern anticipated, predefined changes to the product itself. Where performance is affected by factors other than a change to the product itself, this likely falls out of the scope of a PCCP and could be addressed through PMS or local governance arrangements.
A substantial part of the discussion focused on version control and rollback, with participants noting that an AI system may need to revert to a previous model version, an alternative AI-enabled version, or, in some cases, a non-AI state, and that where rollback is triggered by a safety issue this should feed directly into post-market surveillance (PMS) processes. In this context, a PCCP should clearly define the conditions under which rollback would occur as it relates to changes in the PCCP, the approach to versioning, how version retirement would be managed, and how users would be informed where appropriate. Participants also highlighted that rollback to non-AI or pre-digital states may be difficult to achieve in practice because healthcare services often depend on electronic systems, staffing models and workflows evolve over time, and reverting may be operationally burdensome or unrealistic.

It was further noted that major outages or critical infrastructure failures likely fall outside the scope of a PCCP and are better addressed through incident response and business continuity arrangements rather than planned change control. The discussion also raised the role of deployers in maintaining service continuity, including the tension between strong deployer obligations under DCB-style expectations and the different emphasis of MDR frameworks, which can create uncertainty in practice over where responsibility for continuity controls should sit.

The participants felt that the implications of the discussion are that:

- Generative AI systems can produce outputs that vary across repeated use as a feature of how they work, meaning that conventional validation approaches and standard performance metrics cannot be straightforwardly applied to sufficiently manage the uncertainty of performance, particularly at edge cases

- Human oversight might be assumed to take place equally over the lifetime of a product's deployment, but this may not necessarily be the case. Changes in the quality of human oversight as a category of change requiring management could be considered.

Managing change in AI based tools for rare diseases: Eye2Gene

This session illustrated challenges that arise when AI is applied in clinical contexts characterised by low prevalence, long diagnostic timelines, and rapidly evolving scientific knowledge.

Post-market surveillance under rare disease constraints

For inherited retinal diseases, the patient populations likely to be encountered at any given deployment site are small, genetic testing outcomes that could validate AI predictions may take considerable time to obtain, and the scientific knowledge base against which performance should be assessed is itself evolving as new gene associations are characterised.

Participants agreed that PMS infrastructure should be designed before deployment rather than retrofitted, and that the signals used for monitoring should be defined prospectively rather than identified retrospectively after an adverse event.

Benchmark datasets and their evolution

A significant practical issue discussed by the delegates concerns the use of independent benchmark datasets to evaluate system performance following a change. For a tool expected to expand its scope over time, whether by covering additional genes and variants, serving new patient populations, or being deployed across different imaging hardware, the representativeness of a static benchmark dataset will diminish as deployment context evolves. Participants agreed that benchmark datasets may likely need to be updated alongside the product, but noted that updating a benchmark dataset could itself raise regulatory questions if it alters the agreed upon reference point against which performance is assessed.

Changes to intended use

A central discussion in this session concerned whether the addition of new genes or genetic variants to the system's functionality could constitute a change in intended use, requiring formal reassessment, or could be managed as planned changes within a PCCP. While participants agreed that changes not pre-specified in the originally approved PCCP, or not supported by adequate pre-market evidence, would require reassessment. Discussion also considered whether post-deployment changes could be made to the PCCP itself, subject to re-approval by the Approved Body (AB). Delegates noted that it is important for the evidence standard for demonstrating that a new gene or variant has been added safely and needs to be defined in a way that is proportionate and realistic to the available evidence in rare disease contexts, where sufficient prospective clinical data may be impossible to generate within a reasonable timeframe.

The delegates felt that the implications are:

- Standard assumptions built into the PMS and PCCP framework around an ongoing flow of real-world performance data may not translate in rare disease contexts where patient populations are small, data may be scarce, and updates can be sporadic

- A PCCP may depend on a stable benchmark dataset against which post-change performance can be evaluated for certain changes, but for a tool designed to expand its gene and variant coverage over time, a single, static benchmark likely becomes progressively less representative of what the system is intended to do

- Guidance on how to manage extension of scientific scope (for example, across population, genes or care settings) of a device would provide useful clarity.

Conclusions

The workshop demonstrated that there is substantial appetite for clearer regulatory guidance on managing change across the lifecycle of AI medical devices. Four cross-cutting insights emerged from the workshop.

- Intended use remains the primary regulatory anchor for determining whether a change falls within or outside the types of changes that could be managed within a PCCP, but an intended use statement cannot function effectively for this purpose unless it is drafted with sufficient precision and detail, and interpreted in relation to the full set of contextual factors that shape an AI system's risk profile.

- Effective change management depends on robust monitoring infrastructure and pre-specified thresholds: without pre-specified acceptance criteria, version control arrangements and rollback routes, a PCCP will not offer a sufficiently bounded and auditable framework.

- Not all modifications can be sufficiently evaluated for significance on the basis of the scale of technical modification, and human factors, workflow integration, and the quality of human oversight are important to treat as integral dimensions of significant change assessment rather than secondary considerations due to the impact they can have on the product's use in practice.

- Since AI performance can vary over time and across different healthcare settings, patient populations, workflows, and data conditions, it is important for PCCPs to be informed by real-world use, supported by robust post-market surveillance, and structured so that planned changes are guided by evidence on safety, effectiveness, and clinical relevance rather than assumptions made solely during pre-market development.

Key suggestions from the delegates to the MHRA

The attendees identified the following areas where they felt regulatory guidance would be most valuable:

Clarify that changes to intended purpose and increases in risk class fall outside PCCP scope. Delegates felt that clear guidance is needed to define the boundaries of PCCP use and ensure regulatory consistency, and highlight that PCCPs are intended for anticipated lifecycle changes within the existing purpose and risk profile. Modifications that raise device risk classification should not be managed solely via a PCCP.

Provide clearer expectations on benchmarking datasets and structured testing. It would be helpful for guidance to include information on what constitutes an appropriate benchmark dataset, how often benchmarking should occur, and whether approaches can be standardised or use-case specific.

Clarify how expanding knowledge-based systems can be managed over time. Clear guidance would help explain how scientific validity, clinical performance, and intended purpose should be reassessed as these systems evolve over time, and where the boundary lies between changes that can be included within the PCCP and those that become sufficiently significant to alter the intended purpose and trigger regulatory reassessment. Such guidance should also clarify how ongoing performance monitoring can be used to determine whether a change can be accommodated within the PCCP while ensuring that the device continues to perform safely and effectively without deviating from its intended purpose.

Set clearer expectations for communication of updates. It would be helpful to provide clearer guidance on which categories of change should trigger user communication automatically, particularly where they affect outputs, performance thresholds, interpretation or workflow, and which changes may appropriately be communicated in more passive ways.

Clarify the interaction between PMS and PCCPs. Formal guidance to clearly define how PMS findings should inform the PCCP execution, and highlight when real-world evidence should override planned updates, is important.

Develop a glossary of terms and aligned definitions. A clearer and more accessible glossary of terms would be highly beneficial, since several key concepts are interpreted differently across software, AI and traditional medical device contexts. While terms such as “significant change”, “harm”, “safety”, “performance”, “intended purpose/intended use” and “residual risk” are defined across various standards and guidance, delegates noted that the definitions could be clearer, collated and aligned, with specific examples would reduce ambiguity across context specific nuances.

Develop practical PCCP guidance, templates and educational resources. For example, delegates felt that scenario-based guidance, decision trees, structured examples and flowcharts, worked examples and case studies, tailored educational materials and standardised PCCP templates would provide significant clarity to manufacturers.

International harmonisation. The delegates encouraged that MHRA guidance on significant change and PCCPs be developed with international harmonisation as an explicit objective to prevent unnecessary divergence, enabling manufacturers to navigate multiple regulatory jurisdictions without duplicating submissions unnecessarily.