BMJ Digital Health & AI

Identifying and understanding significant change due to drift when assessing AI models in healthcare: a narrative review

Ylenia Rotalinti,[1,2] Johan Ordish,[1] Xiaoxuan Liu,[3] Ben Glocker,[4] Alastair Denniston,[3] Peter Wright,[1] Christopher Yau,[5] Aditya Kale,[6] David Grainger,[1] Richard Branson,[1] Puja Myles,[1] Allan Tucker[2]

To cite: Rotalinti Y, Ordish J, Liu X, *et al*. Identifying and understanding significant change due to drift when assessing AI models in healthcare: a narrative review. *BMJ Digit Health* 2026;**2**:e000085. doi:10.1136/bmjdh-2026-000085

Received 4 August 2025
Accepted 24 February 2026

[1]Medicines and Healthcare Products Regulatory Agency, London, UK
[2]Brunel University London, London, UK
[3]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
[4]Imperial College London, London, UK
[5]Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[6]University of Birmingham, Birmingham, UK

Correspondence to
Ylenia Rotalinti;
ylenia.rotalinti@mhra.gov.uk

ABSTRACT

Artificial intelligence (AI) as a Medical Device (AIaMD) or medical devices that use AI algorithms—like any other medical device—must meet the requirements of medical device regulation. For regulatory purposes, the most relevant requirement is that the developer must provide evidence that the device performs as intended under normal conditions of use for its entire lifecycle. However, healthcare data are not static and underlying characteristics can change for many reasons (eg, the introduction of new technologies which improve measurement accuracy, changes in population demographics, etc). This 'drift' may lead to a change in performance overall or in certain subgroups in AI models. Models can be updated with new data if significant drift is identified, but in the context of AIaMD, this needs to be done transparently and within a robust regulatory framework. This paper reports on the consensus view of an expert working group hosted by the UK Medicines and Healthcare products Regulatory Agency (MHRA). It aims to highlight the challenges with identifying and assessing significant changes in the performance of a model and understanding the nature of a detected drift to preserve patient safety. We discuss distinct drift subtypes from a statistical perspective and highlight potential causes in the real world that could lead to significant changes to the performance of AI algorithms. We also outline the regulatory challenges associated with risk assessment and the characteristics of drift that are crucial to examine (such as speed and severity) to correctly address interventions and ensure the deployment of safe healthcare products on the market. Finally, we discuss a range of considerations to best identify, risk-assess and intervene for drift when assessing healthcare AI products.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Artificial intelligence (AI) models used in healthcare can experience performance degradation over time due to data drift arising from changes in populations, clinical practice, measurement processes or labelling standards. Existing literature provides statistical definitions and detection methods for different drift types, but guidance on how to interpret drift in relation to clinical risk, regulatory obligations and lifecycle management remains fragmented. Regulatory frameworks increasingly recognise the need for postdeployment monitoring but lack a unified conceptual framework for drift assessment.

WHAT THIS STUDY ADDS

⇒ This review provides a structured taxonomy of drift mechanisms relevant to AI as a Medical Device, distinguishing data-driven and model-driven sources of change and illustrating their real-world clinical causes. It introduces a practical risk assessment framework based on drift velocity, magnitude, cause and impact, offering interpretive guidance to support consistent evaluation across clinical and regulatory contexts. The paper bridges technical definitions of drift with ethical, legal and regulatory considerations, including fairness and data protection.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The proposed framework supports more systematic and proportionate responses to drift in deployed AI systems, informing postmarket surveillance, update strategies and regulatory decision-making. It can guide manufacturers in designing predetermined change control plans and help regulators and healthcare organisations interpret drift beyond purely technical metrics. More broadly, the work encourages a total product lifecycle approach to AI governance that treats drift as an expected and manageable aspect of clinical AI deployment.

INTRODUCTION

Artificial intelligence (AI) as a Medical Device (AIaMD) is a medical device that incorporates AI, specifically machine learning (ML) approaches.[1–3] While they may offer significant advantages in healthcare, AIaMDs present unique characteristics that challenge

existing regulatory frameworks developed for dealing with traditional medical devices.[4]

One challenge is that the underlying characteristics of health data used to train AIaMD are likely to change over time, potentially impacting performance of devices and assumptions underpinning the initial authorisation.[5] Changes in health data could happen for several reasons,[6] for example, due to the launch of new technologies on the market, which improve measurement accuracy or changes in population characteristics (eg, age, diet and genetics) that may be harder to forecast as the COVID-19 pandemic has shown.[7] This phenomenon is known as drift.[8 9]

This paper presents the consensus view from an expert working group (EWG) involving regulatory experts, clinicians, data scientists, industry representatives and health service AI implementation experts, hosted by the UK Medicines and Healthcare products Regulatory Agency (MHRA) to understand approaches to detecting and assessing significant changes in AI models' behaviour as well as the nature of an observed drift and its regulatory implications. We discuss three distinct subtypes of drift from a statistical perspective (covariate, target and concept drift) and highlight potential causes in the real world that could lead to significant shifts in performance requiring actions such as model recalibration or retraining on new data. We also discuss the regulatory perspective on risk assessment and the key features of drift (such as speed and severity) that must be considered to properly address interventions and guarantee the release of safe medical products onto the market.

We further consider how regulatory approaches (such as predetermined change control plans (PCCPs) required under medical device regulations, data protection regulations and the European Union (EU) AI Act) can provide mechanisms to manage drift responsively and fairly. By framing drift not only as a technical issue but also as a regulatory and ethical one, we argue for a total product lifecycle (TPLC) model that enables continuous model monitoring, recalibration and governance across the entire deployment journey.

DRIFT

This section formalises the framework for drift detection, providing consolidated definitions from a statistical perspective and presenting three different subtypes of drift depending on the source of change, along with real-world examples.

Drift from a data science viewpoint

Applications of ML in healthcare typically involve prediction models *trained* on a range of inputs X (or covariates such as height, smoking status), which are associated with a known outcome y (target variable such as the onset of diabetes in 5 years).[10]

Many ML tasks require data samples collected over an extended period of time t with associated timestamps. However, data updates may often be available in blocks of data (ie, batches) at specific intervals where regularity and frequency (eg, annual, monthly) depend on the specific scenario. Here, we assume a batch learning scenario, where blocks of data b (of potentially different sample sizes) are released and processed all at once at a given time. While this is a common paradigm for deployed clinical AI systems, alternative approaches such as online learning or reinforcement learning-based adaptation can also incorporate new data incrementally.[9]

If the assumption that the *statistical distribution of the new data matches that of the original data used to train the AI model* does not hold true, the model may no longer be applicable to the new data; the MHRA EWG agreed that this scenario could be considered under the wider umbrella of 'drift'.

Change can be triggered by different reasons, depending on which factor of the joint probability has shifted. Regulatory systems for AIaMD need to be designed so as to support the recognition, communication and mitigation of this change and of its multiple potential causes. The next section presents a conceptual overview of the three different subtypes of drift identified from a data science perspective and possible events that could lead to the observed drift (table 1 includes the mathematical representations).

Taxonomy of drift subtypes and potential clinical causes within healthcare data

In this work, we use the term *drift* as an overarching concept encompassing changes that may arise from either the data or the model itself. Under this umbrella, covariate drift and target drift are treated as forms of *data drift*, whereas concept drift is classified as a form of *model drift*, as it reflects changes in the model's decision logic or operational definition rather than shifts in input data distributions.

Covariate drift

Covariate drift[11 12] is the most common form of data drift. It happens when the distribution of the covariates $P(X)$ changes. Generally, it occurs when new data (on which a model is assessed after deployment) differ from the training data.

Causes of covariate drift might involve:

1. Change in population characteristics: changes in characteristics of populations (eg, distribution of age, sex, diet, behaviour, ethnicity and genetics) or subpopulations. This can be introduced by both temporal changes and spatial changes, that is, the geographic areas considered by the model.
2. Change in measurement method or data quality: change in how a feature is measured. For instance, if visual acuities are measured on light boxes of lower intensity than standard (caused by the falling over time of bulb brightness), the measure will be subjected to an unexpected change. This may also happen due to routine software updates of a scanner or the introduc-

Table 1 The following table presents a taxonomy of drift, categorising it into three subtypes and providing their mathematical definitions

Drift type	Mathematical definition	Potential metrics to detect change	Potential causes in clinical contexts	Clinical examples
Covariate drift	$P_{b1}(X) \neq P_{b2}(X)$	Distance measures between feature distributions (Kolmogorov-Smirnov test, Kullback-Leibler divergence and Hellinger distance) or feature correlation	Change in population characteristics	A model trained on speech recordings from native English speakers in a single region performs poorly when deployed nationally, where patients present with different accents, dialects and linguistic structures
			Change in measurement method	A mammography scanner software update, altering pixel intensity distributions and causing a threefold increase in recall rates without any model change
			Change in sample selection	A model trained in a controlled research environment performing poorly when applied to a broader, more heterogeneous clinical population
			Changes within interfeature correlation	Age–obesity correlations shifting due to population-level lifestyle or dietary pattern changes
			Changes due to hidden variables	Post-pandemic increases in advanced-stage cancer presentations due to delayed screening programmes, altering the distribution of presenting cases
Target drift	$P_{b1}(y) \neq P_{b2}(y)$	Distance measures between label distributions (Kolmogorov-Smirnov test, Kullback-Leibler divergence and Hellinger distance)	Changes in target outcome prevalence	COVID-19 pandemic altering pneumonia or mortality rates, thereby affecting models trained on pre-pandemic outcome distributions
			Changes in label assignment	Sepsis prediction models trained using one institutional definition performed poorly when deployed in hospitals using modified sepsis criteria or multistage severity labels
Concept drift	$P_{b1}(y\|X) \neq P_{b2}(y\|X)$	Performance metrics (accuracy, sensitivity, specificity, precision, recall, F1-score and AUC)*	Changes in annotation guidelines	Revision of diagnostic criteria (eg, HbA1c thresholds or updated clinical guidelines), altering how diseases are annotated
			Changes in annotation meaning	Transition from ICD-9 to ICD-10 coding causes differences in disease definitions that change model outputs
			Changes in operating point	Modifying the decision threshold for a deployed model (eg, sensitivity/specificity balance), changing how outputs are interpreted clinically
			Inconsistencies in the labelling process	Use of semiautomated tools like CheXpert introducing inconsistency in labels such as 'No Finding' due to ambiguous radiology report phrasing

Assuming a batch learning scenario, data drift occurs when changes are detected between two blocks of data, $b1$ and $b2$, due to a shift in one or more factors of the joint probability $P(X,y)$. Specifically, covariate drift is identified when there is a change in the distribution of the covariates X, target drift is triggered by a change in the distribution of the target variable y and concept drift represents a change in the relationship between X and y. To detect these changes, it is possible to use various quantitative metrics. Examples of such metrics are provided in the table. Each subtype of drift is then divided into subcategories based on potential clinical causes.
*While these metrics can signal the presence of drift, they do not uniquely identify concept drift, as changes in performance may also be driven by covariate or target drift.

tion of a new technology that improves the measurement resolution.

3. Change in sample selection: change in which individuals are included in the data used to develop, validate or operate the model. This may be caused by dissimilarities between a controlled training environment and a target population, leading the model to not generalise correctly.
4. Change within interfeature correlation: change in the relationship among observed features. For example, when considering age and obesity as two inputs to a model, the relationship between them might change over time due to lifestyle changes or new dietary patterns.[13]
5. Changes due to hidden variables: changes due to hidden variables not measured directly, but that have a direct impact on other observed features. For instance, after an effective disease awareness campaign, more patients may proactively seek a clinical diagnosis even with few symptoms, increasing the reporting and discovery of new cases. Another example is changes in the performance of early COVID-19 diagnostic models, which did not include COVID-19 vaccination as a variable following the introduction of the vaccine.

As an additional real-world example, consider a hypothetical natural language processing (NLP) model designed to support general practitioners in detecting early signs of Alzheimer's disease. The model was trained

on recordings from native English speakers in a specific geographic region, where speech patterns were relatively uniform. However, when deployed nationally, the model encountered a much broader variety of accents, dialects and linguistic structures. This geographic and linguistic diversity introduced covariate drift by altering the distribution of speech features, particularly due to changes in population characteristics and sample selection. Consequently, the model began generating false positives—flagging patients as cognitively impaired when they were not. In some cases, clinicians over-relied on the model's output, even when standard screening tests (eg, memory or spatial assessments) indicated otherwise.

Another example of how drift can occur due to hidden variables, measurement and changes in data quality was observed during the COVID-19 pandemic when many routine cancer screening programmes were temporarily suspended. When services resumed, the prevalence of detected cancers rose sharply due to delays in diagnosis and a backlog of missed appointments. This resulted in a significant shift in the underlying distribution of cases encountered by predictive models—an instance of covariate drift driven by hidden variables and changes in recorded disease prevalence over time. For example, studies[14] reported increases in advanced-stage diagnoses in post-pandemic screenings, necessitating recalibration of decision thresholds to maintain model accuracy in a changed clinical environment.

While the above example of drift was partly due to changes in measurement policy (ie, suspension of screening), covariate drift can also occur without any changes in population but instead due to alterations in measurement methods. In a retrospective evaluation of an AI system for breast cancer screening, de Vries *et al*[15] found that a subtle software update to the mammography equipment led to significant performance degradation. Although the AI model remained unchanged, the software update affected image characteristics, resulting in a threefold increase in recall rates. This form of drift requires software version recalibration of decision thresholds to restore acceptable performance levels. It serves as a reminder that even seemingly minor technical changes (if not documented and addressed) can profoundly affect downstream AI performance.

Target drift

If the distribution $P(y)$ of the variable we are trying to predict (ie, the target, outcome or predicted variable) changes, a target drift is encountered.[16]

Some of the causes for target drift might be:

1. Changes to target outcome prevalence: change in the prevalence of the predicted variable. This may be linked to changes in the sample selection, which not only leads to covariate drift but also simultaneously contributes to overall drift. For example, rising obesity rates have increased the incidence of cardiovascular disease, while declining smoking prevalence has reduced the burden of tobacco-related illness. Although rare global events such as the COVID-19 pandemic can cause sudden and substantial shifts in outcome prevalence (eg, marked increases in pneumonia rates),[17] these represent exceptional rather than routine drivers of target drift.
2. Change in label assignment: change in how the variable we are predicting is labelled/coded. It refers to variation in how an existing definition is applied to instances, resulting in shifts in the distribution of the target variable. For instance, some hospitals may use multiple categories denoting disease stage or severity, rather than using a binary label denoting presence or absence of disease. An example is sepsis, which is variably defined and has been shown to generalise poorly when models trained using data with one definition are subsequently implemented in settings using different definitions.[18]

Concept drift

When the relationship between the inputs and the decision that has been given based on that data changes, we define this as concept drift.[19] This change will typically degrade the model's performance; however, it is important to note that performance deterioration cannot be attributed solely to concept drift. Shifts in covariates or in outcome prevalence and labelling (covariate and target drift) can produce similar performance patterns.

Some of the causes might be:

1. Intentional change in the 'ground truth' by an externally defined gold standard (eg, recognised by experts).
   - Change in annotation guidelines: an example is a change to clinical guidelines used to diagnose a disease because a new diagnostic criterion has been introduced or previous thresholds/criteria have been modified. For example, this may occur when medical data taxonomies are expanded to explicitly include non-binary genders and intersex identities, requiring updates to how demographic or clinical categories are annotated.
   - Change in annotation meaning: change of the terminology due to a new coding system or a new policy. For instance, a model developed with diagnoses defined using an International Classification of Diseases 9th revision (ICD-9) code may yield different results when deployed in institutions that have adopted ICD-10 because of differences in definitions.
2. Intentional change in a model's internal definition of truth (carried out by the manufacturer on advice by clinical experts) caused by:
   - Change in operating point: changes in the threshold used to discriminate between positive or negative examples, that is, sensitivity/specificity thresholds. This would have a large impact on the model performance and, therefore, on the product safety.
3. Unintentional deviation from 'ground truth' over time or space due to:

- Inconsistencies in the labelling process: change in how datasets are annotated. Datasets may be exposed to inconsistencies in the labelling process carried out by human annotators due to (potentially unconscious) expert bias, differences in judgement, noise, lack of concentration and subjectivity. Additionally, the use of semiautomated labelling tools, such as the CheXpert labeller (which applies NLP to extract pathology labels from radiology reports), can introduce inconsistencies in annotation. This is particularly evident in labels such as 'No Finding', where ambiguity or imprecision in report language can result in inaccurate ground truth data, thereby contributing to concept drift if not properly accounted for.[20]

Different coexisting causes might contribute simultaneously to the overall drift since some causes may potentially trigger more than one subtype of drift. For example, the vaccination programme introduced during the COVID-19 pandemic would potentially trigger both concept drift (the relationship between COVID-19 incidence and mortality would be different for a vaccinated population than an unvaccinated one) and target drift (a reduction in disease prevalence due to vaccination). This aspect is undoubtedly a challenge when trying to draw a direct causal relationship to identify the nature of the observed drift.

In practice, clinicians and manufacturers can follow a staged diagnostic process: first, assess whether performance deterioration is present, which may indicate concept drift or severe target drift; second, investigate changes in population characteristics or measurement quality to determine whether covariate drift is contributing and finally, review labelling practices or gold-standard definitions to detect shifts in target definitions. This staged approach does not attribute causality deterministically but enables prioritisation of likely root causes and supports proportionate intervention. In operational settings, this process often requires collaboration between clinical domain experts, data scientists and regulatory specialists to ensure that observed drift is interpreted correctly and addressed in a manner aligned with clinical risk and regulatory obligations.

Recent work[21] further highlights that even when individual sources of drift can be identified, determining the extent to which each contributes to overall model degradation is substantially more challenging. Even when the underlying causes can be identified, attributing proportions of overall performance degradation to individual mechanisms (such as concept, covariate or target drift) remains an open challenge. For example, in sepsis prediction, early interventions triggered by model deployment may simultaneously alter patient physiology (covariate drift), clinician ordering behaviour (target drift) and the mapping from features to labels (concept drift). These intertwined effects underscore that drift assessment must go beyond distributional distance measures and performance monitoring to account for more complex, system-level changes.

Different types of drift also have distinct implications for clinical decision-making and risk management. For instance, when treatment decisions are based on predefined risk thresholds, covariate drift may not degrade model calibration sufficiently to necessitate model refitting. The model may continue to rank patients correctly, and its discrimination performance can remain stable. However, because covariate drift alters the distribution of input features, it can substantially change the proportion of patients whose predicted risk exceeds the treatment threshold. In such cases, the *volume* of individuals receiving treatment may shift without any change in model parameters. By contrast, concept drift affects the underlying relationship between inputs and outcomes, degrading clinical validity and requiring model updating or recalibration.

The advantage of defining a clear taxonomy by including also a mathematical definition of such events, as presented in table 1, is that we can formally cover all the types of drifts from a mathematical perspective. In addition, we can design specific metrics (eg, Kullback-Leibler distance[22] or classic performance metrics[23] to detect, quantify and distinguish different forms of drift. From a regulatory standpoint, this acknowledgement is essential to address the mitigation strategy and subsequent actions to an occurred drift.

The next section describes crucial factors to classify a change within an authorised AI health product into one of four categories, from lowest to highest risk, to reflect the risk associated with the clinical situation and device use.

Drift risk assessment

The following aspects must be considered when assessing the risk associated with an observed drift, particularly in relation to patient safety and broader regulatory obligations.

Drift velocity

This term refers to the rate and direction at which change occurs in a model's input data, output labels or the relationships between them. Understanding the speed of drift (whether it is gradual or abrupt) can provide insight into the underlying causes. For example, gradual drift may arise from slow-moving factors such as demographic changes, whereas abrupt drift could result from sudden events like the implementation of a new clinical guideline or the introduction of a novel diagnostic device. To aid interpretation, time-based reference points can be helpful: a gradual drift might unfold over several years (eg, 5–10 years), while an abrupt drift could manifest within days or weeks. The direction of change is also critical: a drift may indicate a deterioration in model performance, raising safety concerns, or an unexpected improvement that still warrants scrutiny. Importantly, the implications of directionality depend on which aspects of model performance are affected. For instance, a drift that reduces sensitivity may raise immediate safety

concerns by increasing missed diagnoses, whereas a drift that reduces specificity increases false positives, which can trigger unnecessary follow-up tests, anxiety and avoidable treatments. As a result, the choice of evaluation metrics (and whether they are applied in short-term performance monitoring or longer-term outcome assessment) must align with the anticipated impact of the drift and the clinical role of the AIaMD. Evaluating both speed and direction helps determine whether intervention is needed and how urgently it should be implemented.

Drift magnitude

Quantifying the severity of a drift is essential to determine when to trigger further investigation. Severity should be interpreted through prespecified thresholds linked to clinical concern, ranging from early warnings to critical failure states. In practice, the magnitude of drift does not have a single universal unit; rather, it may be expressed through different quantitative indicators depending on the model and the clinical task. For instance, it could reflect changes in a model performance metric (such as a reduction in area under curve (AUC)), a shift in calibration error, or the number or proportion of patients whose clinical management would differ due to altered predictions. These thresholds, however, may need to be dynamically updated over time as the consequences of the drift become clearer, especially in the context of evolving technologies. For instance, there may be a temporal lag between the occurrence of a drift and its observable clinical consequences, making it necessary for the framework to support ongoing refinement of risk thresholds. Importantly, magnitude alone may not be sufficient to determine clinical impact. There may be an absolute threshold below which increasing drift has negligible or no clinical effect, and hence does not warrant action. Conversely, even minor shifts might be significant in high-stakes contexts. Therefore, magnitude must always be interpreted in context—in light of the intended use of the AIaMD, its safety-critical role and the setting in which it operates.

Drift causes

Understanding possible causes of drift is essential to guide an appropriate response. These may include changes in population, practice, measurement processes or labelling quality (see previous section). Clinical and technical expert input should be involved to identify the most plausible root causes.

Drift impact (regulatory actions/manufacturer duties)

Estimate the level of potential harm related to different levels of failing performance. Drift impact may be considered across several domains, including:

A. Clinical impact: consequences for diagnostic accuracy, treatment decisions and patient safety.
B. Regulatory impact: compliance obligations arising under medical device legislation.
C. Data protection impact: fairness and lawful processing concerns under the UK General Data Protection Regulation (GDPR).
D. Equity and ethical impact: potential disparate effects across protected groups.
E. Operational impact (implications for workflow or maintenance obligations).

To inform the action needed in response to an observed drift, we must consider its nature. A sudden, impactful drift would require fast action in terms of understanding and managing the interventions, although the specific corrective action may not necessarily be extreme. For example, a sudden, high-magnitude drift affecting a diagnostic or triage system may warrant immediate mitigation actions such as temporary suspension of deployment, urgent investigation or notification to relevant authorities. In contrast, a gradual, low-magnitude drift observed in a screening or decision-support system may be appropriately managed through enhanced monitoring, documentation and planned recalibration within routine maintenance cycles. Depending on the nature of the drift, there will be different implications. From a regulatory perspective, this includes not only medical device regulators but may also involve data protection and AI regulators. Even where the drift poses minimal clinical safety concern, it may still raise legal or ethical issues, such as disadvantaging particular groups in access to healthcare services due to shifts affecting protected characteristics under the Equality Act. In particular, from a data protection standpoint, drift in an AI system that processes personal data (especially health data, which is classed as a special category under the UK GDPR) raises concerns around the fairness principle. According to the UK Information Commissioner's Office (ICO),[24] fairness means using personal data in ways that individuals would reasonably expect and that do not result in unjustified adverse effects. If drift leads to inaccurate outputs or unintended bias, this may render the use of AI unfair, particularly if it has unduly detrimental, unexpected or misleading consequences.

From the medical device regulatory perspective, specifically, it is useful to align with existing risk management methodologies to quantify drift-related risk. Some types of change are inherently riskier than others. For instance, a sudden, high-magnitude drift due to a new policy or clinical guideline—rated medium-high or high using a risk matrix—would likely trigger model updates regardless of root cause. In contrast, a slow drift caused by demographic changes might present a lower risk and, if performance stays within predefined limits, may not require immediate retraining.

This principle is illustrated in figure 1, which shows how different drift trajectories (sudden, gradual and low-level) interact with predefined thresholds to inform regulatory decision-making. Sudden drifts can quickly exceed action or catastrophic thresholds, necessitating prompt intervention, whereas slower drifts may stay within acceptable bounds for longer. The figure also highlights

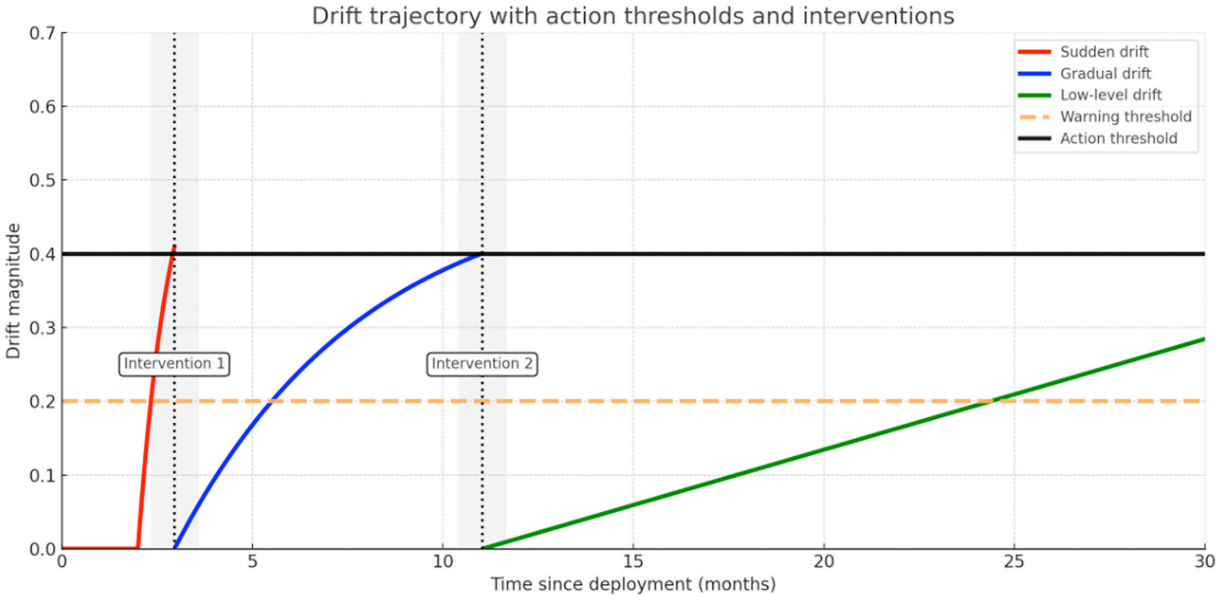

Figure 1 Drift trajectories over time relative to regulatory thresholds. The figure represents a system experiencing three sequential forms of drift: sudden drift (red), gradual drift (blue) and low-level drift (green). Warning (dashed orange) and action (solid black) thresholds indicate increasing levels of concern. Drift magnitude is shown on a normalised scale to illustrate relative change rather than a specific metric; the concrete definition of this magnitude may vary depending on the drift detection method and clinical context. Interventions are triggered when drift magnitude crosses the action threshold, as shown by the vertical dotted lines. Each intervention resets the drift magnitude to zero before the next phase begins. Shaded regions highlight the intervention windows.

the importance of aligning intervention timing with the velocity and magnitude of the drift, emphasising that delayed action, particularly in the context of high-velocity drifts, can increase patient safety risks. These trajectories serve as practical tools for interpreting the clinical and regulatory relevance of performance changes over time and support a risk-based approach to model update decisions. In practice, the specific numerical thresholds used to trigger warnings or actions will depend on the clinical context and the acceptable level of risk. For example, a screening system may tolerate higher thresholds for false positives with primarily operational consequences, whereas diagnostic or triage systems may require much lower thresholds because even small increases in false negatives can pose significant patient safety risks.

Ultimately, the risk analysis should be clinically, ethically and legally driven. Seemingly large changes in performance may not matter clinically, while minor changes may have outsized consequences—particularly in relation to fairness, data protection or equality. Drift must always be interpreted in light of the AIaMD's intended purpose, the deployment pathway and its broader implications for patient safety, legal compliance and individual rights.

To ensure a robust and well-rounded assessment, input should be sought not only from clinical and technical experts, but also from domain specialists in policy, law and regulation, particularly where drift may intersect with legal duties under data protection or equality legislation, or carry policy-level consequences for the health system or population groups.

Regulatory perspective

In 2021, the US Food and Drug Administration (FDA), Health Canada and the UK's MHRA identified 10 guiding principles for Good Machine Learning Practice,[25] emphasising the importance of monitoring deployed models for performance and managing retraining risks. These principles are particularly critical for mitigating risks associated with model drift, which can significantly impact patient safety if not properly addressed.

Regulatory bodies have begun to outline pathways for managing such drifts. For instance, the FDA's recent emphasis on PCCPs[26] offers a structured approach to managing anticipated changes in AI models. A PCCP includes specific plans for model modifications, methods for validating these changes and protocols for reverting updates that negatively impact performance. This approach ensures that regulators and manufacturers are aligned in maintaining high standards of safety and effectiveness throughout the product life cycle.

Importantly, emerging regulatory guidance places primary responsibility for postdeployment monitoring and management of drift on the manufacturer or software provider. The PCCPs explicitly require manufacturers to anticipate, monitor and manage model changes throughout deployment, including defining appropriate update strategies and validation procedures. However, the expected intensity and frequency of monitoring are inherently context-dependent. Models trained on large, stable historical datasets or deployed in relatively static populations may require minimal updates,

whereas systems trained on limited data or operating in rapidly changing clinical environments may necessitate more frequent monitoring and adaptation. Regulatory oversight bodies do not typically perform continuous monitoring themselves but instead enforce compliance through postmarket surveillance obligations, audits and accountability mechanisms, ensuring that appropriate monitoring processes are in place and proportionate to risk.

However, current regulatory frameworks may not fully address all risks associated with drift. The rapid pace of AI and ML developments can exceed the traditional, more static regulatory pathways. A principle-based approach to regulation, as highlighted by various stakeholders, could offer more flexibility, allowing for adaptive responses to new risks while maintaining patient safety. This could involve incorporating real-time data monitoring, risk-based assessments and evidence generation throughout the product life cycle, ensuring that AI models are continually optimised for accuracy and fairness.

In this context, data protection authorities are also emerging as key regulatory actors. In March 2023, the UK ICO released updated guidance on AI and data protection,[27] applicable to all uses of AI that process personal data, including AIaMD. This guidance underscores that drift is not only a clinical safety risk but also a data protection risk, particularly in relation to fairness, accuracy and compliance with the UK GDPR. For example, the ICO notes that statistical accuracy must be maintained postdeployment, and warns that distributional or concept drift can lead to unjustified adverse effects if systems are not properly monitored and adjusted.

Adding further complexity, the EU AI Act,[28] which nominally entered into force in August 2024, introduces a new crosscutting regulatory layer. Though many of its requirements are being phased in through August 2027, the Act establishes a comprehensive framework for AI governance across the EU and beyond. Under the EU AI Act, AI-as-a-medical-device is likely to be classified as a high-risk AI system, given that medical devices are covered under EU harmonisation legislation and must undergo conformity assessment prior to being placed on the market or put into service. This classification carries a number of obligations for both providers and deployers, many of which directly address the types of risks introduced by drift.

By adopting a TPLC perspective,[29] regulators can ensure that drift is not only detected but appropriately managed. The TPLC framework spans the premarket phase (design, data curation and validation), market authorisation and postmarket performance monitoring; for AIaMDs, drift considerations are particularly relevant during the latter two phases, where proactive surveillance, impact assessment and managed updates ensure continued safety and effectiveness. This would involve integrating continuous monitoring mechanisms, postmarket surveillance and stakeholder collaboration into the regulatory process, creating a dynamic and responsive framework that can adapt to changes as they occur.[30]

By systematically categorising different types of drift and proposing specific risk assessment methodologies, this work offers valuable insights that can directly enhance the formulation and execution of PCCPs. The detailed taxonomy of drift types, combined with real-world examples and risk-based approaches, equips regulators, developers and stakeholders with the tools needed to anticipate and manage changes throughout a device's lifecycle. By incorporating these insights into PCCP plans, manufacturers can ensure their AI models remain safe, effective and compliant with regulatory standards.

CONCLUSION

This paper provides a structured overview of different types of drift that can affect AIaMD products across time and settings. Using a statistical framework, we defined and classified drift into three subtypes—covariate, target and concept drift—each with distinct causes and implications. For each subtype, we outlined clinical scenarios and real-world events that could trigger a significant change in model performance, potentially leading to patient harm or regulatory non-compliance.

We proposed a practical risk assessment framework based on drift velocity, magnitude, cause and impact, and introduced interpretive guidance for terms like 'gradual', 'abrupt' and 'severe' to support consistent application in regulatory and clinical settings. We also highlighted the importance of understanding inherent versus residual risk, and the role of prespecified thresholds in determining when to act. Crucially, we stressed that risk evaluation should not only rely on clinical and technical perspectives but must also include expert input from legal, regulatory and policy domains—particularly when drift implicates fairness, equality or data protection concerns.

From a regulatory standpoint, the paper contextualises drift within evolving global frameworks. It discusses mechanisms such as the PCCPs in the context of medical device regulations, the UK's data protection obligations under GDPR (as clarified in recent ICO guidance) and the EU Artificial Intelligence Act, which imposes specific lifecycle and postmarket monitoring duties on high-risk AI systems, including AIaMDs.

Collectively, these insights support a TPLC approach to AI regulation that recognises drift as an expected and manageable phenomenon, rather than an exception. By combining robust monitoring with proactive planning and stakeholder collaboration, developers and regulators can ensure that AI models remain safe, fair and effective throughout their use in healthcare.

Despite the progress in defining and regulating drift, several open research challenges remain. First, there is a need for clinically reliable drift detection methods that can operate under real-world constraints such as sparse labels or delayed outcomes. Second, adaptation strategies that update models safely and transparently (without compromising

traceability, regulatory approvals or postmarket commitments) are still immature. Third, methods capable of attributing drift to underlying causes, especially when multiple drift types co-occur, remain limited and require interdisciplinary input to avoid misinterpretation. Finally, further work is needed to integrate technical drift monitoring with governance frameworks that account for fairness, equity and data protection concerns. Addressing these gaps will be essential to support the safe and scalable deployment of AIaMDs in dynamic healthcare environments.

Acknowledgements This project has been made possible by a grant from the Regulators' Pioneer Fund launched by the Department for Business, Energy and Industrial Strategy (BEIS). The fund enables UK regulators and local authorities to help create a UK regulatory environment that unleashes innovation and makes the UK the best place to start and grow a business. The authors would like to extend their sincere gratitude to Chris Russell, Dean Bodenham, Francesca Edelmann, Guido Fumagalli, Hugh Harvey, Lena Cordie-Bancroft and Marina Evangelou, who participated in the workshops hosted by the Medicines and Healthcare products Regulatory Agency (MHRA) and generously shared their insights and expertise. Their contributions were invaluable in shaping the discussions and enhancing the quality of this research.

Contributors YR is the guarantor of this work. YR led the drafting of the manuscript and contributed to the conception and design of the paper. JO, XL, BG, AD, PW, CY, AK, DG, RB, PM and AT contributed to the development of key concepts, participated in expert discussions and provided substantial revisions to the manuscript. All authors participated in workshops hosted by the MHRA and contributed critical intellectual content. All authors reviewed and approved the final version of the manuscript.

Funding This work was supported by the UK Regulators' Pioneer Fund under grant G2-SCH-2021-09-8644.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, conduct, reporting or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCE

1 Software and AI as a medical device change programme. Available: https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-changeprogramme [Accessed 1 Jul 2023].
2 The regulation of artificial intelligence as a medical device. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1120503/RHC_regulation_of_AI_as_a_Medical_Device_report.pdf [Accessed 1 Jul 2023].
3 Machine learning AI in medical devices. Available: https://www.ethos.co.im/wp-content/uploads/2020/11/MACHINE-LEARNING-AI-IN-MEDICAL-DEVICES-ADAPTING-REGULATORY-FRAMEWORKS-AND-STANDARDS-TO-ENSURE-SAFETY-AND-PERFORMANCE-2020-AAMI-and-BSI.pdf [Accessed 1 Jul 2023].
4 The medical devices regulations 2002. Available: https://www.legislation.gov.uk/uksi/2002/618/contents/made [Accessed 1 Jul 2023].
5 Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD).
6 Finlayson SG, Subbaswamy A, Singh K, *et al*. The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med* 2021;385:283–6.
7 Sáez C, Romero N, Conejero JA, *et al*. Potential limitations in COVID-19 machine learning due to data source variability: A case study in the nCov2019 dataset. *J Am Med Inform Assoc* 2021;28:360–4.
8 Gama J, Žliobaitė I, Bifet A, *et al*. A survey on concept drift adaptation. *ACM Comput Surv* 2014;46:1–37.
9 Lu J, Liu A, Dong F, *et al*. Learning under Concept Drift: A Review. *IEEE Trans Knowl Data Eng* 2018;31:1.
10 Shailaja K, Seetharamulu B, Jabbar MA. Machine learning in healthcare: a review. 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA); IEEE, 910–4. Coimbatore.
11 Ditzler G, Roveri M, Alippi C, *et al*. Learning in Nonstationary Environments: A Survey. *IEEE Comput Intell Mag* 2015;10:12–25.
12 Sugiyama M, Kawanabe M. *Machine learning in non-stationary environments: introduction to covariate shift adaptation*. MIT press, 2012.
13 Management and prevention of obesity and its complications in children and adolescents. Available: https://www.mja.com.au/journal/2005/182/3/3-management-and-prevention-obesity-and-its-complications-children-and [Accessed 1 Jul 2023].
14 Trimarco V, Izzo R, Pacella D, *et al*. The COVID-19 pandemic increased the incidence of newly diagnosed cancers: evidence from a large cohort study in Southern Italy. *BMC Med* 2025;23:399.
15 de Vries CF, Colosimo SJ, Staff RT, *et al*. Impact of Different Mammography Systems on Artificial Intelligence Performance in Breast Cancer Screening. *Radiol Artif Intell* 2023;5:e220146.
16 Zhang K, Schölkopf B, Muandet K, *et al*. Domain adaptation under target and conditional shift. InInternational Conference on Machine Learning 2013 May 26; PMLR, 819–27.
17 Flynn D, Moloney E, Bhattarai N, *et al*. COVID-19 pandemic in the United Kingdom. *Health Policy Technol* 2020;9:673–91.
18 Wong A, Otles E, Donnelly JP, *et al*. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Intern Med* 2021;181:1065–70.
19 Webb GI, Hyde R, Cao H, *et al*. Characterizing concept drift. *Data Min Knowl Disc* 2016;30:964–94.
20 Irvin J, Rajpurkar P, Ko M, *et al*. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *AAAI* 2019;33:590–7.
21 Ansari S, Baur B, Singh K, *et al*. Challenges in the Postmarket Surveillance of Clinical Prediction Models. *NEJM AI* 2025;2:AIp2401116.
22 Lovric M. Kullback-leibler divergence. In: *International encyclopedia of statistical science*. Berlin, Heidelberg: Springer, 2011: 720–2.
23 Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. *Pattern Recognit Lett* 2009;30:27–38.
24 Information Commissioner's Office. How do we ensure fairness in AI? Available: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidanceon-ai-and-data-protection/how-do-we-ensure-fairness-in-ai/ [Accessed 1 Jul 2023].
25 US Food and Drug Administration. Good machine learning practice for medical device development: guiding principles. Available: https://www.fda.gov/media/153486/download [Accessed 1 Jul 2023].
26 Medicines and Healthcare products Regulatory Agency. Predetermined change control plans for machine learning-enabled medical devices: guiding principles. 2023. Available: https://www.gov.uk/government/publications/predetermined-change-control-plans-formachine-learning-enabled-medical-devices-guiding-principles [Accessed 1 Jul 2023].
27 Information Commissioner's Office. Guidance on ai and data protection. 2023. Available: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ [Accessed 1 Jul 2023].
28 Future of Life Institute. AI Act Explorer. 2024. Available: https://artificialintelligenceact.eu/ai-act-explorer/ [Accessed 1 Jul 2023].
29 US Food and Drug Administration. Total product life cycle for medical devices. Available: https://www.fda.gov/about-fda/cdrh-transparency/total-product-life-cycle-medical-devices [Accessed 1 Jul 2023].
30 Feng J, Phillips RV, Malenica I, *et al*. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digit Med* 2022;5:66.