INTERNATIONAL JOURNAL OF PHARMACEUTICAL SCIENCES
[ISSN: 0975-4725; CODEN(USA): IJPS00]
Journal Homepage: https://www.ijpsjournal.com

OPEN ACCESS

Review Article

The Evolution Of Materiovigilance For Adaptive Artificial Intelligence And Software As A Medical Device (Samd): Algorithmic Drift, Post-Market Safety, And Regulatory Frameworks

Kishore D.[1], Dhinesh Selvaraju[2], Prithiviraj A.[3]*

[1,2,3]*Department of Pharmacy Practice, J.K.K.Nattraja College of Pharmacy, Kumarapalayam, Namakkal, TamilNadu, 638183, India*

ARTICLE INFO

Published: 1 Aug. 2026
Keywords:
materiovigilance; software as a medical device; algorithmic drift; algorithmovigilance; post-market surveillance; artificial intelligence regulation.
DOI:
10.5281/zenodo.21738842

ABSTRACT

Artificial intelligence (AI) and machine learning (ML) are now embedded in hundreds of regulated medical products, yet the frameworks that govern their post-market safety were designed for physical devices whose performance is fixed at manufacture. This mini-review examines a structural tension at the centre of contemporary medical-device safety: adaptive AI and Software as a Medical Device (SaMD) do not behave like static instruments, because their real-world performance degrades through algorithmic drift and, in continuously learning systems, may change by design. Synthesising evidence from original clinical, methodological, and regulatory-science studies indexed in PubMed, we show that drift is frequently silent—discrimination can appear stable while calibration deteriorates—so that miscalibrated risk estimates reach the bedside without any perceptible malfunction. Analyses of regulatory databases reveal that the pre-market evidence for cleared AI devices is thin and narrow, and that post-market adverse events and recalls are dominated by software- and change-related causes, precisely the failure modes that spontaneous adverse-event reporting is least equipped to detect. Materiovigilance, the medical-device analogue of pharmacovigilance, therefore faces an inflection point. We argue that its evolution requires convergence with algorithmovigilance: active, quantitative, equity-aware, and continuous surveillance of deployed model performance, embedded within health systems and harmonised across the divergent regulatory architectures of the United States, European Union, and India. We map the resulting knowledge gaps, compare emerging lifecycle-oriented regulatory instruments, and outline a research agenda for post-market performance monitoring of adaptive medical AI.

***Corresponding Author:** Prithiviraj A
Address: *Department of Pharmacy Practice, J.K.K.Nattraja College of Pharmacy, Kumarapalayam, Namakkal, TamilNadu, 638183, India*
Email ✉: prithivi0812@gmail.com

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

INTRODUCTION

The clinical translation of artificial intelligence (AI) and machine learning (ML) has moved from proof-of-concept to routine regulatory clearance. A comparative analysis of regulatory databases identified 222 AI/ML-based medical devices cleared in the United States and 240 in Europe between 2015 and 2020, with radiology dominating and few products classified as high-risk.[1] The trajectory has since steepened: by mid-2023, 691 AI/ML-enabled devices had received clearance from the US Food and Drug Administration (FDA),[2] and cumulative authorisations subsequently surpassed 870.[3] This expansion has outpaced the conceptual foundations of medical-device safety surveillance, which were built for tools whose behaviour is fixed once they leave the factory.

Post-market surveillance of medical devices—termed materiovigilance—is the coordinated identification, reporting, and analysis of untoward events associated with device use, and it operates as the device-domain counterpart of pharmacovigilance.[4] National programmes such as the Materiovigilance Programme of India (MvPI), launched in 2015, embody the prevailing model: a spontaneous, event-driven reporting system oriented toward discrete, perceptible failures such as implant fracture or stent malfunction.[5] That model rests on an assumption that adaptive software silently violates—that a device which performed acceptably at approval will continue to do so, absent a physical defect, throughout its service life.

For AI-based SaMD, performance is not a fixed property established at authorisation. It decays through algorithmic drift as the statistical relationship between inputs and outcomes shifts over time, even when the underlying code is unchanged. Longitudinal analyses of clinical prediction models have shown that discrimination can remain stable for years while calibration steadily deteriorates, producing systematically biased risk estimates.[6] Performance can also collapse abruptly when the care environment is perturbed, as occurred with models spanning the onset of the COVID-19 pandemic.[7] These are not theoretical concerns: a proprietary sepsis-prediction model implemented at hundreds of US hospitals discriminated poorly and was poorly calibrated on external validation, while generating a heavy burden of alerts.[8]

The challenge is compounded by adaptive systems designed to learn continuously after deployment, which can change their own behaviour in situ and therefore resist the "one device, one approval" paradigm.[9] Regulators have responded with lifecycle-oriented instruments, most prominently the FDA's AI/ML Action Plan and the concept of a predetermined change control plan that pre-authorises specified future modifications.[10] Yet post-market data indicate where the real hazards lie: adverse events and recalls of AI/ML devices are dominated by software- and design-change causes,[2,3] the very failure modes that event-driven materiovigilance is structurally least able to detect. In this review we argue that the evolution of materiovigilance for adaptive AI requires its convergence with algorithmovigilance—the systematic, ongoing, quantitative surveillance of AI-driven care for effectiveness and equity (Fig. 1).[11] We first dissect the mechanisms of algorithmic drift; we then appraise the post-market evidence base, the sociotechnical determinants of real-world performance, and the divergent regulatory architectures now emerging; and we finally

define the knowledge gaps, clinical implications, and research priorities that follow.

2. METHODOLOGY

This is a critical mini-review rather than a systematic review, and it does not follow PRISMA reporting, which is designed for the quantitative synthesis of homogeneous studies; the heterogeneity of the relevant literature—spanning clinical trials, methodological studies, regulatory-database analyses, and health-policy scholarship—precludes meaningful meta-analysis. PubMed was searched as the sole database, consistent with a biomedical scope, for records published between January 2018 and July 2026, with seminal earlier work included where indispensable.

Search terms combined controlled vocabulary and free-text expressions for the core constructs, including "software as a medical device", "artificial intelligence" or "machine learning", "algorithmic drift", "dataset shift", "calibration drift", "concept drift", "post-market surveillance", "materiovigilance", "algorithmovigilance", "predetermined change control plan", and named regulatory frameworks. Priority was given to original research—randomised controlled trials, prospective and retrospective cohort studies, cross-sectional analyses of regulatory databases, and high-impact methodological studies—over narrative reviews, which were consulted only for definitional or historical context. Records were screened by title and abstract for direct relevance to the post-market safety of adaptive AI/SaMD; non-English articles, conference abstracts without peer-reviewed data, and studies unrelated to clinical deployment or its governance were excluded.

For every citation, bibliographic metadata—authors, title, journal, year, volume, issue, and pagination—were retrieved from and verified against the PubMed record before inclusion, and each source was checked to ensure that it supported the specific statement for which it is cited. Evidence was synthesised narratively around the review's central thesis. The principal limitations of this approach are its restriction to a single database and to English-language literature, and the fact that much regulatory guidance on adaptive AI resides in agency documents that are not indexed in PubMed and are therefore discussed as context rather than cited as primary evidence.

3. THE ANATOMY OF ALGORITHMIC DRIFT

3.1 A taxonomy of a family of failures

Algorithmic drift is not a single phenomenon but a family of distinct mechanisms with different signatures and clinical consequences. Covariate or data drift denotes a change in the distribution of model inputs—case mix, coding practices, instrumentation—whereas concept drift denotes a change in the relationship between inputs and the outcome itself. Either can degrade a model, but they degrade it differently, and the distinction is decision-relevant because it dictates whether recalibration or full refitting is the appropriate remedy (Table 3).

3.2 Calibration decay as the dominant, insidious mode

The most instructive empirical characterisation comes from a decade-long analysis of acute kidney injury models developed on Veterans Affairs admissions and validated across nine subsequent years.[6] Discrimination was preserved across all seven modelling methods,

yet calibration declined as the models increasingly over-predicted risk; the magnitude of over-prediction tracked changes in the underlying event rate, whereas shifts in predictor–outcome associations produced method-dependent patterns of miscalibration. Notably, flexible learners such as random forests and neural networks retained calibration better than parametric regression, indicating that drift susceptibility is partly a property of model architecture rather than of the data alone. The central lesson is that the most common failure mode is not a loss of rank-ordering but a loss of calibration—an error that a discrimination-centric evaluation will not reveal.

That calibration decay generalises beyond a single outcome. A lifelong-learning framework evaluated across colorectal, lung, breast, and prostate cancer cohorts likewise attributed performance deterioration to distributional change and demonstrated that jointly monitoring model performance and the input distribution can both detect and explain drift.[12] Drift is not always gradual: using the COVID-19 pandemic as an extreme perturbation, an emergency-admission model retained reasonable discrimination (area under the curve 0.86 before versus 0.83 during the pandemic) while explainability methods exposed abrupt shifts in feature behaviour that signalled emergent, previously unseen risk.[7] Together these studies delineate a spectrum from slow, silent calibration decay to sudden regime change, both of which can occur without any code modification and neither of which announces itself to the user.

This mechanistic picture explains why algorithmic drift is a materiovigilance blind spot. A drifting model emits no alarm, produces no visible malfunction, and continues to return plausible outputs; the harm is statistical and diffuse rather than discrete and attributable. Conventional adverse-event reporting depends on a clinician recognising an event and connecting it to a device, but a subtly miscalibrated probability that nudges a treatment decision satisfies neither condition. The surveillance question is therefore not whether a device failed, but whether it is still performing as validated—a question that can be answered only by continuous quantitative measurement (Fig. 2).

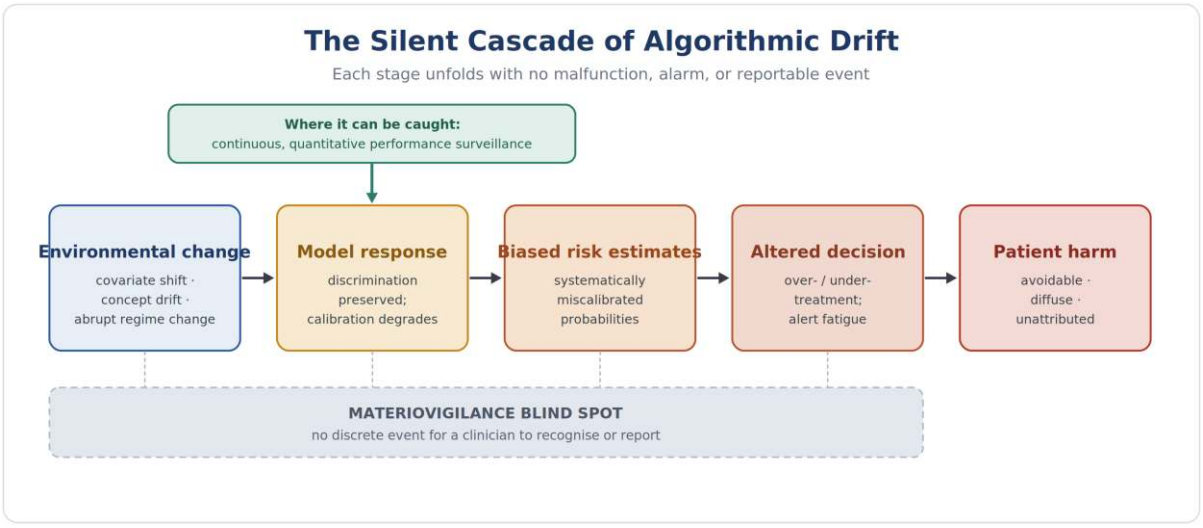

Fig. 1. The silent cascade of algorithmic drift and the materiovigilance blind spot it exposes.

The pathway progresses from environmental change (covariate or concept shift, or abrupt regime change) through preserved discrimination but degrading calibration, to systematically biased risk estimates, altered clinical decisions, and patient harm; each step occurs without a perceptible malfunction or reportable event.

Table 1. Taxonomy of algorithmic drift and corresponding post-market surveillance and mitigation approaches.

Drift type	Mechanism	Typical signature	Detection metric	Mitigation	Example
Covariate (data) drift	Change in input distribution (case mix, coding, instruments)	Calibration decay; discrimination often preserved	Input-distribution monitoring; calibration-in-the-large	Intercept recalibration	Davis 2017 [6]
Concept drift	Change in input–outcome relationship	Method-dependent miscalibration; possible discrimination loss	Predictor–outcome monitoring; recalibration slope	Model refitting	Davis 2017/2019 [6,26]
Abrupt / regime shift	Sudden environmental change (e.g., pandemic)	Rapid performance change; new feature behaviour	Explainability tracking; drift alarms	Rapid retraining; temporary suspension	Duckworth 2021 [7]
Specification / label bias	Flawed optimisation target or proxy	Systematic subgroup disparity	Stratified subgroup performance audit	Re-specification of target	Obermeyer 2019 [19]
Sociotechnical (human–AI) drift	Change in user reliance or behaviour	Divergence of system vs model performance	Human-factors and workflow monitoring	Training; trust calibration; interface change	Dratsch 2023 [18]

Detection metrics and mitigations are indicative rather than exhaustive; in practice, multiple drift types co-occur and require combined surveillance.

4. FROM APPROVAL TO BEDSIDE: THE POST-MARKET EVIDENCE GAP

The gap between the evidence available at authorisation and the reality of deployment is the central safety problem for medical AI (Table 1), and its archetype is the external validation of a widely implemented proprietary sepsis model.[8] Across 27,697 patients, the model achieved a hospitalisation-level area under the curve of only

0.63, failed to identify roughly two-thirds of septic patients, and generated alerts on nearly one in five hospitalisations, imposing substantial alert fatigue. That a tool with such performance had been adopted at hundreds of institutions without adequate independent evaluation illustrates how weakly post-market performance is scrutinised once a product is in routine use.

Systematic appraisals confirm that this is not an isolated case. An analysis of the evidence underpinning FDA approvals found that most AI devices were evaluated retrospectively, frequently at a single site and without prospective or multi-site testing.[13] A systematic review of randomised trials of ML interventions identified only 41 such trials, none fully adherent to the CONSORT-AI reporting standard; half were single-site, more than a third were in gastrointestinal endoscopy, and participants from under-represented minority groups were scarce.[14] The pre-market evidence base is thus not merely thin but narrow, concentrated in a few specialties and populations, which limits any inference about how devices will perform across the heterogeneous settings in which they are ultimately deployed.

Rigorous real-world evaluation is nonetheless achievable and informative. A prospective, multi-site study of a machine-learning sepsis early-warning system monitoring 590,736 patients found that timely provider confirmation of alerts was associated with reduced in-hospital mortality, organ dysfunction, and length of stay among the 6,877 patients identified before antibiotic initiation.[15] In endoscopy, an open randomised trial showed that a real-time detection system raised the adenoma detection rate from 20.3% to 29.1%, chiefly by finding diminutive lesions,[16] but a subsequent double-blind, sham-controlled trial by the same group yielded a smaller effect (34% versus 28%; odds ratio 1.36, 95% CI 1.03–1.79).[17] The attenuation under blinding is itself an important finding: it shows that the measured benefit of an AI device is entangled with operator behaviour, complicating both efficacy estimation and any later attempt to attribute a change in post-market performance to the algorithm rather than to its users.

The clearest window onto post-market safety comes from analyses that link regulatory databases. A cross-sectional study of 691 FDA-cleared AI/ML devices found that only 1.6% reported randomised-trial data and 7.7% reported prospective data, while demographic information was absent in 95.5%; over the study period, 489 adverse events were recorded across 36 devices—predominantly malfunctions, with 30 injuries and one death—and 40 devices were recalled a total of 113 times, chiefly for software problems.[2] A complementary root-cause analysis of nearly three decades of AI/ML device recalls found that design and software-design factors accounted for approximately half of recalls and that software changes and control-related changes contributed substantially.[3] These are the decisive data for the present argument: the empirical signature of post-market harm in medical AI is software- and change-driven—exactly the class of failure that event-based device vigilance detects last and least well.

Table 2. Summary of landmark and representative studies informing the post-market safety of adaptive AI/SaMD.

Study	Design / data source	Focus	Key finding	Ref
Davis et al, 2017	Longitudinal cohort, US Veterans Affairs (2003–2012)	Calibration drift (AKI)	Discrimination stable but calibration decayed (over-prediction); drift linked to case-mix/event-rate change; magnitude architecture-dependent.	6
Chi et al, 2022	Multi-cohort (4 cancers)	Drift + lifelong updating	Deterioration attributable to distributional change; joint monitoring of performance and inputs detects and explains drift.	12
Duckworth et al, 2021	ED admissions, COVID-19	Abrupt data/concept drift	Explainability exposed sudden feature shifts; AUC 0.86→0.83 across pandemic onset.	7
Wong et al, 2021	Retrospective external validation (n=27,697)	Deployed sepsis model	AUC 0.63; missed ~67% of sepsis; alerts on ~18% of hospitalisations (alert fatigue).	8
Adams et al, 2022	Prospective, multi-site (n=590,736)	Deployed ML sepsis alert	Timely alert confirmation associated with lower mortality, organ failure, and length of stay.	15
Wang et al, 2019	Open RCT (n=1,058)	CADe colonoscopy	Adenoma detection rate 29.1% vs 20.3%, chiefly diminutive lesions.	16
Wang et al, 2020	Double-blind, sham-controlled RCT	CADe colonoscopy	ADR 34% vs 28% (OR 1.36); effect attenuated relative to open trial.	17
Dratsch et al, 2023	Controlled reader study (27 radiologists)	Automation bias	Incorrect AI suggestions degraded reader performance; worst among inexperienced readers.	18
Obermeyer et al, 2019	Retrospective algorithm audit	Specification/racial bias	Cost-as-proxy produced racial bias; correction	19

			raised Black patients flagged 17.7%→46.5%.	
Lin et al, 2025	Cross-sectional, FDA databases (n=691)	Pre/post-market reporting & safety	1.6% reported RCTs; 7.7% prospective; 489 adverse events/36 devices; 40 devices recalled 113 times.	2
Chen et al, 2025	Recall root-cause analysis (27 years)	AI/ML device recalls	Design/software-design ~50% of recalls; software and control changes substantial.	3
Plana et al, 2022	Systematic review (41 RCTs)	Evidence quality	No trial fully CONSORT-AI compliant; single-site and endoscopy-heavy; low minority inclusion.	14

AKI, acute kidney injury; AUC, area under the receiver-operating-characteristic curve; ADR, adenoma detection rate; CADe, computer-aided detection; ML, machine learning; RCT, randomised controlled trial; SaMD, Software as a Medical Device.

5. HUMAN FACTORS AND THE SOCIOTECHNICAL SURFACE OF DRIFT

Deployed AI does not act on patients directly; it acts through clinicians, and its real-world performance is a property of the human–machine system rather than of the model in isolation. In a controlled study, 27 radiologists interpreting mammograms with a purported AI aid were measurably swayed by incorrect BI-RADS suggestions, with the largest degradation among less experienced readers.[18] Automation bias of this kind means that a model's errors are not merely passed through but can be amplified by the humans who rely on it, and that the safety of a device therefore depends on factors—user experience, workflow, trust calibration—that lie outside the algorithm and outside the scope of any purely technical validation.

A related and more insidious failure arises when the target a model optimises is a flawed proxy for the clinical goal. A widely used population-health algorithm was shown to exhibit substantial racial bias because it predicted health-care cost rather than illness; at equal risk scores, Black patients were considerably sicker, and correcting the proxy would have raised the proportion of Black patients flagged for additional care from 17.7% to 46.5%.[19] Such specification bias is not a transient drift but a latent, systematic distortion that persists undetected unless performance is monitored separately within demographic subgroups. It follows that post-market surveillance of medical AI must be sociotechnical and stratified: it must observe the human–AI system as deployed and must disaggregate performance by subgroup, rather than tracking a single pooled accuracy metric.

6. REGULATORY ARCHITECTURES FOR A MOVING TARGET

Regulating an artefact that can change confronts agencies built to certify fixed products. The most trenchant framing of this problem argues that regulators must shift from a product view to a system view, evaluating not only the algorithm but the clinical system within which it learns and acts.[9] The distinction between "locked" algorithms, which are frozen at deployment, and "adaptive" algorithms, which update from new data, is central to every current framework, because the two pose opposite risks: locking forecloses improvement while remaining vulnerable to environmental drift, whereas adaptation permits correction but admits the possibility of uncontrolled or self-reinforcing change (Table 2).

The FDA's approach has coalesced around total-product-lifecycle oversight. Its AI/ML Action Plan and the associated predetermined change control plan—comprising pre-specified performance boundaries and an algorithm change protocol—seek to authorise anticipated modifications in advance while shifting assurance toward the post-market phase.[10] Earlier experiments, such as the Digital Health Software Precertification pilot, explicitly elevated real-world performance monitoring as a pillar of oversight,[20] and complementary regulatory-science scholarship has argued that causal-inference methods are needed to establish that adaptive systems remain safe and effective as they change.[21] The unresolved difficulty is operational: pre-authorising change presupposes a monitoring apparatus capable of detecting when a system has drifted beyond its sanctioned envelope, and that apparatus is not yet standardised.

The governance of these rules is not a neutral technical exercise. An analysis of public comments submitted on the FDA's proposed framework for modifications to AI/ML-based SaMD found that a majority came from parties with financial ties to industry, that such ties were rarely disclosed, and that the overwhelming majority of submissions cited no scientific evidence at all.[22] This raises a legitimate concern that the standards governing adaptive-AI safety may be shaped disproportionately by interested parties, reinforcing the case for independent, evidence-based post-market surveillance rather than reliance on manufacturer self-attestation.

The European Union has taken a more prescriptive, risk-based route. The 2024 AI Act layers obligations for high-risk systems, including many medical devices, on top of the existing Medical Device Regulation and In Vitro Diagnostic Regulation, creating both new duties and potential conflicts around classification, conformity assessment, and post-market monitoring.[23] Comparative analysis across the United States, European Union, and the Republic of Korea shows a broadly convergent emphasis on transparency, bias mitigation, and continuous post-market monitoring, but divergent means: the EU imposes stricter ex-ante oversight, the United States favours flexible pathways that accommodate continuous learning, and Korea leans on real-world data for validation.[24] Against these maturing frameworks, spontaneous-reporting device-vigilance programmes represent an earlier developmental stage, and none of the major systems has yet fully operationalised the continuous performance surveillance that all of them increasingly presuppose.

Table 3. Comparison of major regulatory frameworks for adaptive AI/SaMD.

Dimension	United States (FDA)	European Union	India (MvPI / CDSCO)
Primary instrument	AI/ML Action Plan; total-product-lifecycle approach	Medical Device Regulation + AI Act (2024)	Medical Device Rules 2017; Materiovigilance Programme
Mechanism for change in adaptive AI	Predetermined change control plan (pre-specifications + algorithm change protocol)	Ex-ante conformity assessment; substantial-modification rules; high-risk obligations	Not specifically defined for adaptive software
Regulatory posture	Flexible, lifecycle; accommodates continuous learning	Prescriptive, risk-based; stricter ex-ante oversight	Emerging; oriented to physical-device safety
Post-market requirement	Real-world performance monitoring (emphasised, not standardised)	Post-market surveillance and post-market clinical follow-up obligations	Spontaneous adverse-event reporting
Maturity for algorithmic drift	Conceptually advanced; operationally incomplete	Formally strong; MDR/AI-Act interface unsettled	Early stage; software-performance pathway absent
Key gap	Standardised drift-detection triggers	Harmonisation of AI Act with MDR/IVDR	Continuous performance-surveillance capacity
Supporting refs	10, 20, 21	23, 24	4, 5

CDSCO, Central Drugs Standard Control Organisation; FDA, US Food and Drug Administration; IVDR, In Vitro Diagnostic Regulation; MDR, Medical Device Regulation; MvPI, Materiovigilance Programme of India.

7. MATERIOVIGILANCE AT THE INFLECTION POINT

The Indian experience concretely illustrates both the reach and the limits of contemporary materiovigilance. Since its launch in 2015, MvPI has built a national network of monitoring centres coordinated by the Indian Pharmacopoeia Commission and, between 2015 and 2019, received 1,931 adverse-event reports, of which 1,277 were serious and cardiac stents were the single most reported device category.[5] The programme's design—spontaneous reporting of perceptible device failures, supported by stakeholder awareness efforts—is well suited to physical malfunctions but presupposes an identifiable event and a reporter who recognises it.[4]

This is precisely where the model and the technology diverge. Algorithmic drift produces no discrete event to report; it is gradual, statistical, and invisible at the point of care, so it falls into a structural blind spot of any purely spontaneous, event-driven system. A materiovigilance programme could receive a report for a software crash or an egregiously

wrong output, but not for a model whose calibration has quietly degraded such that its risk estimates are now systematically several percentage points too high. Detecting the latter requires a categorically different instrument—continuous, quantitative measurement of deployed performance—rather than the passive receipt of reports.

The emerging response is algorithmovigilance, conceived as the systematic analysis and ongoing monitoring of AI-driven care for both effectiveness and equity.[11] Its operational counterpart is the proposal to establish dedicated hospital units for AI quality assurance and improvement that adapt long-standing statistical process-control tools to the monitoring of deployed models and pair them with disciplined procedures for updating.[25] Updating, however, is not a panacea: methodological work shows that the optimal response to drift ranges from no action, through simple recalibration, to full refitting depending on the type and magnitude of the shift, and that over-eager updating risks overfitting to transient fluctuations.[26] The evolution of materiovigilance therefore entails absorbing these quantitative, lifecycle-oriented practices—transforming a passive reporting registry into an active performance-surveillance system without discarding the institutional infrastructure that national programmes already provide (Fig. 3).

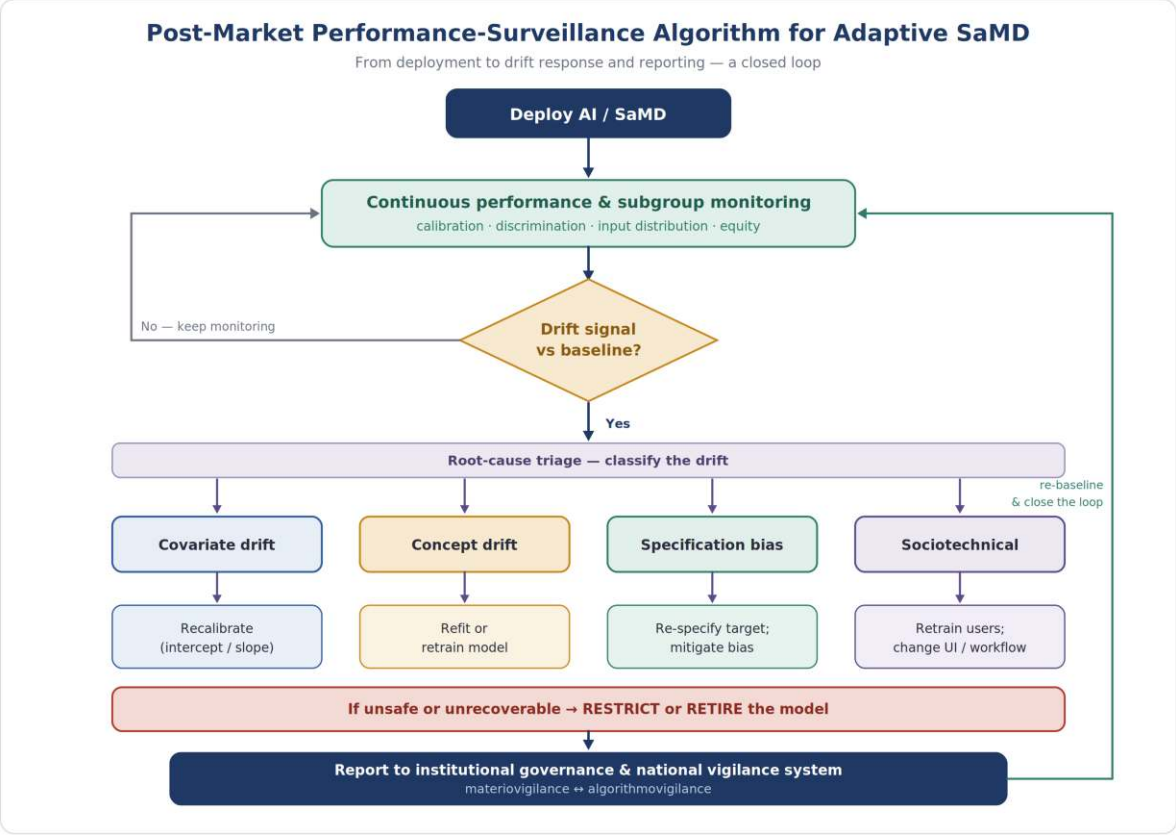

Fig. 3. Proposed post-market performance-surveillance algorithm for adaptive Software as a Medical Device.

The decision flow runs from deployment through continuous performance and subgroup monitoring; a drift signal detected against a (federated) baseline triggers root-cause triage

(covariate versus concept versus specification versus sociotechnical) and a graded response (recalibrate, refit, retrain, restrict, or retire), followed by mandatory reporting to institutional governance and national vigilance systems, closing the loop.

8. CRITICAL SYNTHESIS AND CONTROVERSIES

Several tensions remain unresolved and merit explicit acknowledgement. First, the promise of continuous learning to maintain performance[12] is in direct tension with the assurance burden it creates, because every autonomous update is, in effect, a new device that has not been independently validated,[10] and no consensus exists on who should validate updates or how often. Second, the evidence base on drift is itself of limited external validity: the most rigorous longitudinal characterisations derive from single health systems,[6,26] and multi-institutional, prospective monitoring studies are scarce, so the field's understanding of how drift behaves across diverse settings is incomplete. Third, attribution of harm is genuinely difficult in sociotechnical systems, where model error, automation bias, and workflow interact,[17,18] complicating both causal analysis and any liability framework. These are not reasons for inaction but a map of where methodological and regulatory development is most needed.

9. KNOWLEDGE GAPS

Four gaps are especially consequential for patient safety (Table 4). The first is the absence of validated, standardised triggers for post-market action: there is no accepted threshold of calibration or discrimination decay at which a deployed model should be recalibrated, retrained, or withdrawn, leaving such decisions ad hoc. The second is the paucity of prospective, multi-site, longitudinal surveillance data; because most drift evidence is retrospective and single-institution,[6,12] the natural history of deployed-model performance across health systems is poorly characterised. The third is the near-total neglect of equity in surveillance: demographic data are frequently unreported in device documentation,[2] minority representation in trials is low,[14] and subgroup performance drift is rarely tracked, so bias amplification can proceed undetected.[19] The fourth is definitional and infrastructural: existing reporting standards address early live evaluation[27] and randomised trials[14] but not continuous post-market performance, and national vigilance programmes lack any pathway for software-performance signals.[5] Each gap matters clinically because it corresponds to a class of avoidable harm that current systems cannot presently see.

Table 4. Key knowledge gaps and proposed future directions.

Knowledge gap	Clinical significance	Proposed direction
No standardised drift-action thresholds	Ad hoc, delayed responses to degradation	Validated calibration/discrimination triggers for recalibrate–retrain–retire decisions
Scarce prospective, multi-site longitudinal monitoring	Natural history of deployed performance unknown	Federated surveillance networks; sentinel-style monitoring across sites

Equity blind spot (subgroup drift; missing demographics)	Silent amplification of health disparities	Mandatory stratified, subgroup-level performance reporting
Reporting standards omit continuous performance	Inconsistent, non-comparable post-market data	Extend CONSORT-AI/DECIDE-AI logic to lifecycle surveillance
Materiovigilance lacks a software-performance pathway	Drift invisible to national vigilance systems	Integrate algorithmovigilance into MvPI and pharmacovigilance programmes
Emerging generative / LLM-based SaMD	Novel non-deterministic drift surface	Dedicated monitoring methods for generative outputs

DECIDE-AI, Developmental and Exploratory Clinical Investigations of Decision support systems driven by Artificial Intelligence; LLM, large language model.

10. FUTURE PERSPECTIVES

Progress will depend on building surveillance infrastructure rather than issuing further guidance alone. Concretely, this means embedding continuous performance monitoring—statistical process control adapted to calibration and subgroup performance—within the institutions that deploy models, as proposed for AI quality-improvement units,[25] and federating that monitoring across sites so that drift can be detected against a network baseline rather than a single hospital's history. It also means developing "performance signals" analogous to pharmacovigilance signals: calibration-in-the-small, prediction-distribution shift, and subgroup calibration as routinely reported metrics whose deterioration triggers review.

Maturing the regulatory–operational interface is equally important. Predetermined change control plans should be coupled to mandatory, auditable post-market performance reporting, and the divergent FDA, EU, and MvPI approaches should be harmonised around a common minimum dataset for deployed-model performance,[10,23,24] supported by real-world-data pipelines. Prospective "silent" evaluation before go-live, an extension of the logic of early live-evaluation reporting standards,[27] would allow drift-vulnerable models to be characterised in situ before they influence care. Finally, the rapid arrival of generative and large-language-model-based SaMD introduces a new drift surface—non-deterministic outputs and sensitivity to shifting input distributions and prompts—that current frameworks were not designed to monitor and that will require dedicated methods; this is flagged here as an emerging priority rather than a solved problem. The unifying vision is the embedding of algorithmovigilance within existing materiovigilance and pharmacovigilance programmes, using their institutional reach as the platform for continuous, quantitative oversight (Fig. 4).

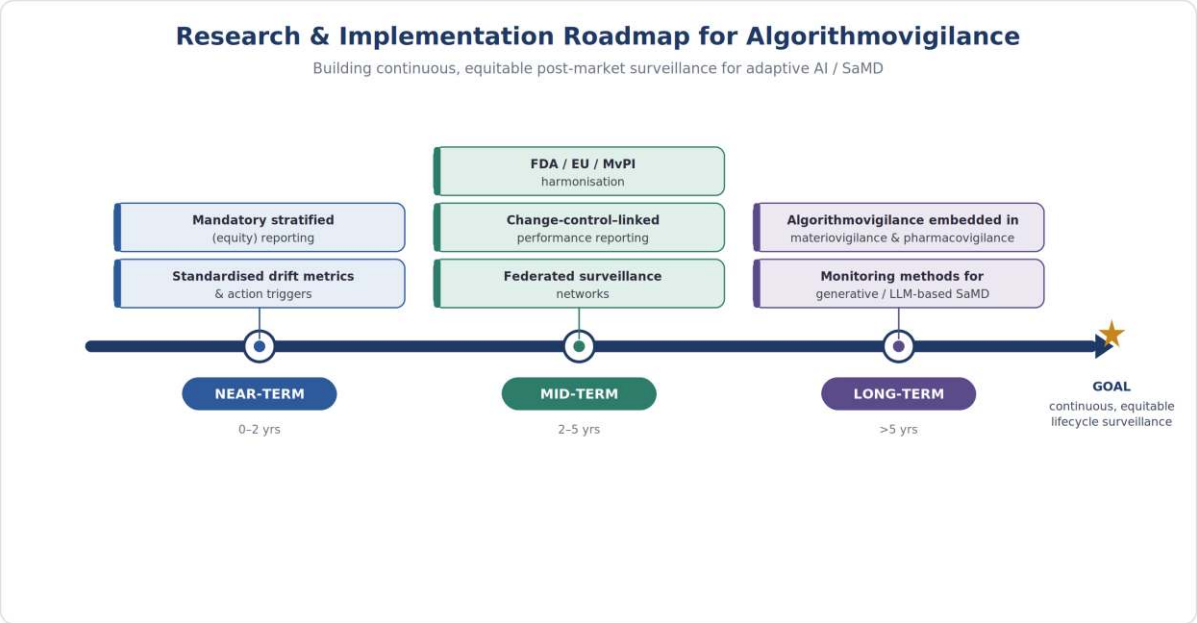

Fig. 4. Research and implementation roadmap for algorithmovigilance.

The timeline links near-term priorities (standardised drift metrics and triggers; stratified reporting), mid-term priorities (federated surveillance networks; change-control-linked mandatory performance reporting; regulatory harmonisation), and long-term priorities (methods for generative/large-language-model-based SaMD; full integration of algorithmovigilance within materiovigilance and pharmacovigilance).

11. CLINICAL IMPLICATIONS

For clinicians, the practical corollary is that AI outputs should be treated as time-varying rather than fixed, and interpreted with awareness of automation bias, particularly by less experienced users.[18] For patients and outcomes, acting on a drifted or miscalibrated risk estimate can translate directly into over- or under-treatment, and poorly calibrated alerting can harm through fatigue as well as through error.[8] For health systems, the implication is organisational: institutions that deploy medical AI require governance and quality-assurance functions capable of monitoring performance over time,[25] and procurement should make post-market performance monitoring a contractual condition linked to the manufacturer's change-control commitments. For policy, national materiovigilance programmes are well positioned to extend their mandate from physical-device malfunctions to software performance,[5] and regulators should require stratified, subgroup-level surveillance to prevent the silent amplification of inequity.[19] Together these measures would shift the locus of safety assurance from a one-time approval toward the continuous stewardship that adaptive technologies demand.

12. CONCLUSION

Adaptive artificial intelligence and Software as a Medical Device violate the assumption on which medical-device safety surveillance was built—that a device which performs acceptably at approval will continue to do so unless it physically fails. Their performance is instead a moving quantity that decays through algorithmic

drift and, in continuously learning systems, changes by design, often silently and without any event that a clinician could recognise or report. The available evidence shows that the safety data accompanying these products at authorisation are thin and narrow, and that the harms observed after deployment are dominated by software- and change-related causes—precisely the failures that spontaneous, event-driven materiovigilance is least able to detect. Regulatory frameworks in the United States, the European Union, and India are converging on lifecycle oversight, yet none has fully operationalised the continuous performance surveillance that such oversight presupposes. The resolution is neither to abandon materiovigilance nor to trust manufacturer self-assessment, but to evolve materiovigilance into an active discipline by fusing it with algorithmovigilance: quantitative, equity-aware, and continuous monitoring of deployed model performance, embedded within health systems and harmonised across jurisdictions. Making the silent visible—measuring whether a model still performs as validated, for every patient it serves—is the defining safety task of medical artificial intelligence, and building the infrastructure to accomplish it is the most urgent priority for the field.

ACKNOWLEDGEMENTS

None

FUNDING

This review received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest relevant to this work**.**

REFERENCES

1. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis. Lancet Digit Health, 2021; 3(3): e195-203.
2. Lin JC, Jain B, Iyer JM, et al. Benefit-risk reporting for FDA-cleared artificial intelligence-enabled medical devices. JAMA Health Forum, 2025; 6(9): e253351.
3. Chen WP, Teng WG, Kuo CB, et al. Regulatory insights from 27 years of artificial intelligence/machine learning-enabled medical device recalls in the United States: implications for future governance. JMIR Med Inform, 2025; 13: e67552.
4. Meher BR. Materiovigilance: an Indian perspective. Perspect Clin Res, 2018; 9(4): 175-8.
5. Shukla S, Gupta M, Pandit S, et al. Implementation of adverse event reporting for medical devices, India. Bull World Health Organ, 2020; 98(3): 206-11.
6. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc, 2017; 24(6): 1052-61.
7. Duckworth C, Chmiel FP, Burns DK, et al. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. Sci Rep, 2021; 11(1): 23017.
8. 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(8): 1065-70.
9. Gerke S, Babic B, Evgeniou T, Cohen IG. The need for a system view to regulate artificial

intelligence/machine learning-based software as medical device. NPJ Digit Med, 2020; 3: 53.
10. Vokinger KN, Feuerriegel S, Kesselheim AS. Continual learning in medical devices: FDA's action plan and beyond. Lancet Digit Health, 2021; 3(6): e337-8.
11. Embi PJ. Algorithmovigilance—advancing methods to analyze and monitor artificial intelligence-driven health care for effectiveness and equity. JAMA Netw Open, 2021; 4(4): e214622.
12. Chi S, Tian Y, Wang F, Zhou T, Jin S, Li J. A novel lifelong machine learning-based method to eliminate calibration drift in clinical prediction models. Artif Intell Med, 2022; 125: 102256.
13. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med, 2021; 27(4): 582-4.
14. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized clinical trials of machine learning interventions in health care: a systematic review. JAMA Netw Open, 2022; 5(9): e2233946.
15. Adams R, Henry KE, Sridharan A, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med, 2022; 28(7): 1455-60.
16. Wang P, Berzin TM, Glissen Brown JR, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut, 2019; 68(10): 1813-9.
17. Wang P, Liu X, Berzin TM, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. Lancet Gastroenterol Hepatol, 2020; 5(4): 343-51.
18. Dratsch T, Chen X, Rezazade Mehrizi M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology, 2023; 307(4): e222176.
19. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 2019; 366(6464): 447-53.
20. King F, Klonoff DC, Ahn D, et al. Diabetes Technology Society report on the FDA Digital Health Software Precertification Program meeting. J Diabetes Sci Technol, 2019; 13(1): 128-39.
21. Stern AD, Price WN. Regulatory oversight, causal inference, and safe and effective health care machine learning. Biostatistics, 2020; 21(2): 363-7.
22. Smith JA, Abhari RE, Hussain Z, Heneghan C, Collins GS, Carr AJ. Industry ties and evidence in public comments on the FDA framework for modifications to artificial intelligence/machine learning-based medical devices: a cross sectional study. BMJ Open, 2020; 10(10): e039969.
23. Kalodanis K, Feretzakis G, Rizomiliotis P, et al. Evaluating the impact of the EU AI Act on medical device regulation. Stud Health Technol Inform, 2025; 323: 40-4.
24. Bottini M, Ryu SJ, Terander AE, et al. The ever-evolving regulatory landscape concerning development and clinical application of machine intelligence: practical consequences for spine artificial intelligence research. Neurospine, 2025; 22(1): 134-43.
25. 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(1): 66.

26. Davis SE, Greevy RA, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. J Am Med Inform Assoc, 2019; 26(12): 1448-57.
27. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ, 2022; 377: e070904.

HOW TO CITE: Kishore D., Dhinesh Selvaraju, Prithiviraj A.*, The Evolution Of Materiovigilance For Adaptive Artificial Intelligence And Software As A Medical Device (Samd): Algorithmic Drift, Post-Market Safety, And Regulatory Frameworks, Int. J. of Pharm. Sci., 2026, Vol 4, Issue 8, 171-187.https://doi.org/10.5281/zenodo.21738842