EXECUTIVE KNOWLEDGE BRIEF

After Clearance

Continuous Monitoring as the Foundation of Clinical AI Oversight

FDA clearance should begin - not end - rigorous evaluation of clinical AI.

Creators
Ank A. Agarwal | Alia N. Sajjadian | Kavita Patel | Roxana Daneshjou

Published July 20, 2026 **Source** NEJM AI editorial
doi.org/10.1056/AIe2600807

WHY THIS MATTERS

The editorial reframes FDA clearance as the start of evidence generation for clinical AI, not the final proof that a model will remain safe across hospitals, patient populations, hardware, workflows, and time.

Prepared July 23, 2026 | Executive analysis for clinical, regulatory, technology, and governance leaders

Contents

A structured guide to the policy argument, monitoring architecture, physician-developer implications, and practical next steps.

Executive Summary

THE STRATEGIC CASE FOR LIFECYCLE OVERSIGHT

Agarwal and colleagues argue that the U.S. regulatory system evaluates clinical AI as though it were static medical hardware, even though model performance is conditional on changing data and deployment environments. Clearance commonly rests on retrospective, curated, and sometimes single-site evidence. Once deployed, an apparently functional model can silently deteriorate because patient demographics, disease prevalence, clinical practice, input devices, acquisition protocols, or software have changed. The model may continue returning confident outputs without an obvious malfunction.

The authors propose a lifecycle approach with two linked reforms. Before clearance, developers should provide multisite external validation, demographic and device transparency, and standardized validation data that the FDA can independently reanalyze. After clearance, a federated surveillance network modeled on FDA Sentinel should continuously monitor deployed tools. Low-cost, label-free checks for input and score-distribution shifts would run routinely; outcome-based evaluation would be escalated when those checks detect a signal. Predefined performance and subgroup thresholds would trigger investigation, mitigation, restrictions, or withdrawal.

The practical significance is larger than regulatory compliance. Continuous monitoring creates an operational safety loop, gives less-resourced hospitals access to shared evaluation capacity, and establishes a contemporaneous record of drift alerts and decisions to maintain, retrain, restrict, or retire a model. For physician-developers, the editorial turns "responsible AI" into an implementable system of telemetry, thresholds, clinical outcome linkage, human adjudication, and corrective action.

EVIDENCE SNAPSHOT

81%	43%	8.1%	17/130
of reviewed deep-learning algorithms performed worse externally	of AI device recalls occurred within 12 months of clearance	of reported device studies were evaluated prospectively	devices reported demographic subgroup performance

Central Argument

Problem

Clinical AI tools can be cleared through pathways designed for static devices on limited retrospective evidence, then used broadly without systematic postmarket evaluation. Because AI behavior depends on the data and environment it encounters, performance can degrade silently after deployment. Current public summaries also provide too little information for hospitals or independent investigators to reproduce validation claims or assess local fit. (pp. 1-2)

Main claim

FDA oversight should require a lifecycle evidence model in which enhanced premarket validation is followed by continuous, real-world postmarket monitoring. Clearance should begin, rather than end, rigorous evaluation. (pp. 2-4)

Supporting reasoning and evidence

- More than 1,000 AI/ML-enabled medical devices have received FDA clearance or authorization. (p. 1)
- In a systematic review of 86 deep-learning algorithms, 81% performed worse on external datasets. (p. 1)
- Forty-three percent of AI device recalls occurred within the first 12 months after clearance. (p. 1)
- Among FDA-authorized AI devices with reported studies, only 8.1% were evaluated prospectively. (p. 1)
- Only 17 of 130 assessed FDA-approved AI devices reported demographic subgroup performance. (p. 2)
- A retinal disease model's error rate increased from 5.5% to 46.6% when used with a different optical coherence tomography scanner. (p. 2)
- Among 1,012 AI/ML device decision summaries, only 1.8% disclosed the exact source of training data and 48.4% reported any performance metric. (p. 2)
- In another analysis, only 3.6% of FDA-cleared devices reported the race or ethnicity of their study populations. (p. 3)

Assumptions

- Distribution shifts can be detected early enough to prevent or limit clinical harm.
- Monitoring metrics and acceptable thresholds can be defined for particular tools and uses.
- Deployment sites can collect sufficiently consistent input, prediction, and outcome data.
- A coordinating body can distinguish harmless distribution change from clinically important degradation.
- The FDA, Congress, CMS, and partner institutions can create sustainable authority and funding for shared surveillance.
- Manufacturers will disclose enough model provenance, known limitations, and failure modes for monitoring to be meaningful.

Limits and unresolved questions

The editorial is a policy argument, not a tested implementation blueprint. It does not specify a universal metric set, threshold-selection method, statistical alerting design, outcome-label latency standard, or technical interoperability architecture. It also leaves open how responsibilities and costs should be divided among developers, health systems, clinicians, regulators, and payers. Multisite surveillance may introduce data harmonization, privacy, governance, and alert-fatigue problems of its own.

ANALYTICAL NOTE

A shift detector is not itself a safety guarantee. It can identify that the production distribution differs from a reference distribution, but only outcome-linked evaluation can establish whether patient-relevant performance has deteriorated. A production oversight program therefore needs two distinct layers: scalable early-warning telemetry and a slower clinical evaluation pathway with human adjudication.

Detailed Knowledge Map

1. Clearance evidence does not guarantee transportability

The authors begin with a mismatch between regulatory evidence and real-world behavior. FDA pathways may accept curated retrospective validation, including evidence from one site, even though a model will later encounter different populations, devices, protocols, and workflows. Prospective trials could improve the evidence base, but no premarket study can fully anticipate behavior in every dynamic deployment environment. The answer is not merely "a better trial"; it is sustained evaluation across the product lifecycle. (p. 1)

EVIDENCE FRAMEWORK

- **Premarket validity:** whether a tool met specified performance requirements on its submitted validation data.
- **Deployment validity:** whether it performs safely and usefully in a particular institution and workflow.
- **Longitudinal validity:** whether that performance remains acceptable as populations, hardware, practices, and prevalence change.

These are related but not interchangeable forms of evidence.

2. Silent failure is the defining operational hazard

The editorial contrasts AI failure with visible hardware malfunction. A failing model can continue to generate a confident answer. Changes in patient mix, disease prevalence, imaging hardware, acquisition protocol, or workflow can alter the relationship between inputs and outcomes without producing an obvious error message. The optical coherence tomography example, where error increased from 5.5% to 46.6% on a different scanner, illustrates an interface failure between model assumptions and the local technical environment. (p. 2)

Related concepts:

- Dataset Shift
- Data Drift
- Concept Drift
- Model Calibration
- Algorithmic Bias
- Silent Failure

3. Transparency is necessary for local oversight

Device developers submit validation results to the FDA, but public decision summaries often omit the source of training data and even basic performance metrics. This prevents independent reanalysis and makes it difficult for a hospital to decide whether the validation population, equipment, and labeling process resemble its own deployment. The authors distinguish standardized FDA access from unrestricted public disclosure: the agency already receives underlying data, but it needs the data in a form that supports reproducible analysis without necessarily exposing proprietary information. (p. 2)

Minimum transparency proposed before clearance:

1. Number and diversity of validation sites.

2. Patient demographics, including race and ethnicity.
3. Device and hardware types.
4. Labeling methodology.
5. Standardized validation data that permits independent reproduction of performance claims.

4. The current and proposed pathways

Figure 1 contrasts two regulatory loops. (p. 2)

Current pathway

1. Retrospective validation, often with a single-site curated dataset.
2. FDA clearance through 510(k) or De Novo; the figure notes that 97% use 510(k) and that prospective data are not required.
3. Full deployment across hospitals.
4. No systematic postmarket evaluation.

Proposed pathway

1. Multisite external validation using two to three diverse sites, with dataset transparency.
2. FDA clearance under enhanced standards, with demographics, devices, and labeling disclosed.
3. Full deployment.
4. Continuous postmarket monitoring.
5. Reevaluation when thresholds are not met.

Reevaluation can also be triggered by software updates, hardware changes, deployment-setting changes, demographic shifts, or evidence of data drift. (pp. 2-3)

5. Existing FDA initiatives do not close the monitoring gap

The Predetermined Change Control Plan lets a developer prespecify future software modifications. The AI/ML-Based Software as a Medical Device Action Plan addresses iterative AI updates. The authors argue that neither requires ongoing real-world performance monitoring across deployment sites. Postmarket surveillance and change control may fall within existing FDA authority, but requiring multisite validation for 510(k) clearance could require special controls, device reclassification, or new authority. A national evaluation network would probably need dedicated Congressional funding. (pp. 2-3)

This distinction matters:

- **Change control** governs anticipated modifications to the product.
- **Performance monitoring** asks whether the product remains safe and effective in actual use.

A model can degrade even when its own code and weights have not changed.

6. A federated evaluation network can extend oversight capacity

The authors propose a network modeled on FDA Sentinel, coordinated centrally by the FDA, that would monitor tools across institutions, independently assess degradation, adjudicate disputed signals, and trigger corrective action. Shared infrastructure would make robust oversight available to hospitals that cannot build a full model-evaluation team. (p. 3)

Potential financing mechanisms include:

- Monitoring obligations as a condition of market access for high-risk devices.
- CMS coverage with evidence development.
- Existing quality-reporting programs.
- Public-private partnerships involving NIH and ARPA-H.

The authors suggest these reforms may cost less than the downstream consequences of tools that fail quietly, although the article does not supply a formal cost analysis. (p. 3)

7. Surveillance should be staged

The operational design uses two levels of monitoring. (p. 3)

Level 1: continuous, low-cost, label-free checks

- Changes in input distributions.
- Changes in model score or output distributions.
- Technical or operational changes that alter the deployment context.

These checks can run on every case and every site.

Level 2: outcome-linked clinical evaluation

- Confirm whether the detected shift changes sensitivity, specificity, false-positive rates, calibration, or subgroup performance.
- Use expert review and outcome labels when Level 1 signals justify the additional cost.

This design concentrates scarce clinical and labeling resources on plausible failures. It also avoids the false premise that every prediction can be immediately paired with a definitive outcome.

8. Thresholds must lead to action

Manufacturers should define minimum sensitivity and specificity, maximum false-positive rates, and acceptable subgroup variation before deployment. Falling below these thresholds or detecting bias should trigger mandatory investigation and a time-bounded mitigation plan. The FDA should be able to require modifications, stronger warnings, use restrictions, or market withdrawal when safety cannot be restored. (p. 3)

The key design principle is that observability without an escalation policy is incomplete. A dashboard does not protect patients unless each signal maps to an owner, review deadline, decision rule, and corrective action.

9. Monitoring records also create accountability

A continuous record of drift alerts and decisions to maintain, retrain, restrict, or retire a model can show when an actor knew or should have known that performance was failing. The authors connect this record to unresolved questions about foreseeability, standard of care, and tort liability for clinical AI harms. (p. 3)

ANALYTICAL NOTE

This makes the monitoring log both a safety artifact and a governance artifact. It should capture not only metrics but also model version, data window, affected population, alert rationale, adjudication, responsible owner, clinical impact assessment, and final disposition.

10. Health systems retain a local duty

Regulatory oversight does not replace local responsibility. Hospitals and community practices should expose known limitations, performance evidence, and failure modes in digital notifications at the point of use. Clinicians need enough context to interpret a result rather than treating it as a context-free answer. This local accountability is difficult when model inputs, training provenance, and known failure modes remain opaque. (pp. 3-4)

Agentic and AI Analysis

This article concerns predictive clinical AI devices rather than autonomous software agents. The monitored system is best classified as a deployed model embedded in a clinical workflow. The proposed oversight layer, however, resembles a tool-using operational workflow:

- **Goal:** Maintain acceptable clinical performance and subgroup equity after deployment.
- **Inputs:** Model inputs, predictions or scores, site and device metadata, software versions, patient demographics, clinical outcomes, and workflow context.
- **Planning:** Primarily prespecified through a monitoring plan, risk tier, alert thresholds, and escalation policy.
- **Tools:** Drift detectors, statistical process control, model and data versioning, clinical data pipelines, outcome adjudication, dashboards, notification systems, and regulatory reporting.
- **Memory:** A longitudinal audit log of deployment context, alerts, investigations, performance, mitigation, and retirement decisions.
- **Execution loop:** Observe -> detect shift -> assess clinical impact -> adjudicate -> mitigate or continue -> document -> reevaluate.
- **Outputs:** Alerts, performance reports, mitigation plans, warnings, restrictions, retraining decisions, or withdrawal.

Where autonomy begins and ends

Automated monitoring can calculate distributions, compare them with baselines, detect threshold violations, and route alerts. It should not independently decide that a tool is clinically safe, change an approved model, suppress a safety signal, or withdraw a device without accountable human and regulatory review.

Required human checkpoints

- Define intended use, reference population, metrics, and thresholds.
- Decide whether a statistical shift has clinical importance.
- Validate outcome labels and investigate confounding workflow changes.
- Assess harm, bias, and subgroup disparities.
- Approve retraining, use restrictions, warnings, or retirement.
- Communicate changes to clinicians and affected operational teams.

Deterministic versus model-dependent elements

Version checks, threshold comparisons, logging, routing, and many statistical tests can be deterministic. Model predictions, drift-detection sensitivity, weak outcome labels, and any AI-assisted root-cause analysis remain probabilistic. Generative AI may help summarize incidents or retrieve documentation, but it should not be the source of truth for safety metrics.

Observability and evaluation needs

A production system needs:

- Stable identifiers for model, software, hardware, site, and data versions.
- Baselines appropriate to each deployment context.
- Stratified performance reporting.
- Monitoring of missingness, input validity, prevalence, calibration, and decision consequences.
- Reliable outcome linkage with known latency.
- Alert quality metrics, including false alarms and missed detections.
- A formal incident response process.
- Tamper-evident audit history and access controls.

Privacy and security

A federated design can reduce movement of patient-level data by computing site-level statistics locally, but privacy is not automatic. The system still needs data minimization, role-based access, secure aggregation, retention rules, and governance for cross-site comparisons. Monitoring infrastructure also becomes a high-value target because it links clinical data, product vulnerabilities, and institutional performance.

Costs and bottlenecks

The most expensive components are likely to be outcome-label acquisition, cross-site data harmonization, expert adjudication, subgroup analysis with small samples, and coordination when a signal spans multiple organizations. Label-free drift checks are comparatively cheap but can create alert fatigue if thresholds are poorly designed.

Requirements for a reliable production workflow

1. A risk-tiered monitoring plan attached to each model's intended use.
2. A machine-readable model and deployment registry.
3. Standardized telemetry and local data-quality checks.
4. Predefined performance and subgroup thresholds.
5. Outcome-linkage and clinical adjudication workflows.
6. Named owners and time limits for every escalation stage.
7. A validated corrective-action process.
8. Periodic testing of the monitoring system itself.
9. Clear reporting interfaces for clinicians, institutional governance, manufacturers, and regulators.

Physician-Developer and Doctors Who Code Relevance

Clinical friction addressed

The editorial addresses a specific safety gap: a clinician or hospital may use a cleared model without knowing that its local scanner, patient population, or workflow differs from the environment in which performance was established. Clearance answers a market-access question; it does not automatically answer the bedside question, "Does this tool still work here, for these patients, today?"

Workflow change

Continuous monitoring changes clinical AI from a one-time procurement and validation event into an operated service. Every deployed model needs telemetry, an accountable owner, regular review, an incident pathway, and an exit plan. This is closer to site reliability engineering than to installing conventional hardware.

Human clinical judgment

Clinical judgment remains essential when interpreting alerts, establishing patient-relevant outcomes, deciding whether a change is harmful, evaluating subgroup disparities, and restricting or retiring a model. Clinicians also need point-of-use communication about known limitations without being overwhelmed by generic warnings.

What a physician-developer would build

- A deployment registry linking model version, site, device, intended use, and owner.
- A monitoring specification with metrics, cohorts, baselines, and thresholds.
- Data pipelines connecting predictions to outcomes.
- Stratified dashboards that expose uncertainty and sample size.
- A safety-event workflow with adjudication and mitigation.
- A clinician-facing explanation of local performance and known failure modes.
- Tests that simulate hardware, population, prevalence, and workflow shifts.

Safety, evidence, interoperability, privacy, and adoption

Safety depends on detecting failure before it becomes routine care. Evidence must be longitudinal and local enough to capture deployment conditions. Interoperability requires consistent representation of model versions, observations, devices, predictions, outcomes, and provenance. Privacy constraints favor federated or distributed evaluation but complicate auditing. Adoption will depend on making monitoring actionable for clinical teams rather than creating another passive dashboard.

Doctors Who Code potential

The source strongly supports a clinically grounded Doctors Who Code article. Its best framing is not a broad warning about AI regulation; it is a concrete engineering argument: every clinical model needs a production safety loop, and physician-developers are well positioned to translate clinical harm into measurable thresholds, escalation rules, and retirement criteria.

Practical Applications

Use now

- Treat FDA clearance as necessary but insufficient during procurement; request multisite, subgroup, device-specific, and prospective evidence.
- Add model version, intended use, hardware dependencies, known failure modes, and accountable owner to the local AI inventory.
- Establish local baseline distributions and subgroup performance before full rollout.
- Put known limitations and local performance context where clinicians encounter the model output.
- Document every decision made after a drift or performance alert.

Test

- Run a retrospective "scanner swap" or device-stratified analysis to identify hardware-dependent performance.
- Compare input and score-distribution monitoring with delayed outcome-based performance to measure which alerts are actually predictive.
- Shadow-deploy a model at a second site before allowing it to influence care.
- Test whether clinicians understand and appropriately respond to point-of-use limitation notices.
- Simulate a demographic or prevalence shift and measure detection time, escalation time, and patient exposure.

Build

- A clinical AI observability starter kit that combines a model registry, data checks, subgroup metrics, drift alerts, and an incident log.
- A machine-readable monitoring plan template for each clinical model.
- A federated evaluation prototype that returns aggregate performance without centralizing patient records.
- A "model safety case" page that brings evidence, limitations, deployment context, monitoring status, and ownership into one view.
- An automated evidence packet for governance committee and regulator review.

Research

- Which label-free signals best predict clinically meaningful degradation?
- How should thresholds adapt to risk, prevalence, sample size, and subgroup uncertainty?
- What monitoring cadence minimizes harm without producing alert fatigue?
- Which clinical outcomes are sufficiently timely and reliable for surveillance?
- How should liability be allocated when developers, hospitals, or clinicians fail to act on an alert?
- What common data model can represent clinical AI telemetry across vendors and sites?

Avoid

- Assuming that regulatory clearance proves local or durable performance.
- Treating generic drift detection as proof of clinical harm or safety.
- Monitoring only aggregate accuracy while ignoring subgroups and calibration.
- Deploying a dashboard without an owner, escalation deadline, and corrective-action policy.
- Retraining automatically on production data without validation and change control.
- Using opaque products whose provenance and known failure modes cannot support meaningful oversight.

Content Opportunities

1. Clearance Is the Beginning of the Clinical AI Safety Case

- **Channel:** Doctors Who Code
- **Central claim:** A cleared model still needs local and longitudinal evidence.
- **Opening:** A retinal model moves to a different scanner and its error rate rises from 5.5% to 46.6%, while still producing confident answers.
- **Audience:** Clinicians, hospital AI leaders, physician-developers, and health technology buyers.
- **Evidence needed:** The cited scanner study, FDA pathway requirements, and one real-world deployment governance example.
- **Format:** Clinically grounded essay with a lifecycle diagram.

2. Build the Monitor Before You Deploy the Model

- **Channel:** Technical tutorial / Doctors Who Code
- **Central claim:** Monitoring is part of the clinical product, not an optional dashboard added after launch.
- **Opening:** Start with a production incident where no system error occurred, but the patient distribution had shifted.
- **Audience:** Physician-developers and ML engineers working in health care.
- **Evidence needed:** A sample dataset, drift-detector behavior, outcome-linked evaluation, and alert-performance results.
- **Format:** Tutorial with code, architecture diagram, and incident runbook.

3. Why Data Drift Is Not the Same as Clinical Failure

- **Channel:** AI education
- **Central claim:** Drift is an early-warning signal; patient-relevant degradation requires outcome evidence and clinical interpretation.
- **Opening:** Two dashboards show the same distribution shift, but only one corresponds to worse sensitivity.
- **Audience:** Clinical informaticians, quality leaders, and developers.
- **Evidence needed:** Technical definitions, illustrative simulations, and examples of benign and harmful shifts.
- **Format:** Explainer with a two-layer monitoring diagram.

4. The Missing Model Incident Log

- **Channel:** Physician-developer / technical commentary
- **Central claim:** A durable record of alerts, decisions, and mitigation is both a safety tool and an accountability mechanism.
- **Opening:** Ask what the institution could prove six months after a model began underperforming.
- **Audience:** Hospital governance, risk, compliance, informatics, and engineering teams.
- **Evidence needed:** Current incident-response standards, medico-legal analysis, and example audit fields.
- **Format:** Article plus downloadable template.

5. A Federated Sentinel Network for Clinical AI

- **Channel:** Clinical technology commentary
- **Central claim:** Shared surveillance can give smaller hospitals access to evaluation capacity without requiring every site to build a full AI safety team.
- **Opening:** Contrast a large academic medical center with a rural hospital deploying the same device but having very different monitoring resources.
- **Audience:** Regulators, payers, health-system leaders, and policy researchers.
- **Evidence needed:** FDA Sentinel operating model, privacy architecture, governance, cost estimates, and rural implementation constraints.
- **Format:** Policy analysis with a network architecture diagram.

6. The AI Procurement Questions Hidden Inside This Editorial

- **Channel:** Doctors Who Code practical guide
- **Central claim:** Procurement should evaluate transportability, observability, and exit criteria, not merely headline accuracy and FDA status.
- **Opening:** A committee receives an FDA-cleared product deck that omits training sites, subgroup results, and hardware dependencies.
- **Audience:** Clinical AI committees, CIOs, CMIOs, and department leaders.
- **Evidence needed:** Example vendor documentation, governance standards, and interviews with procurement stakeholders.
- **Format:** Checklist-driven article or presentation.

Key Terms and Entities

- **510(k):** FDA premarket pathway based on substantial equivalence to a legally marketed predicate device.
- **De Novo:** FDA pathway for certain novel low-to-moderate-risk devices without a predicate.
- **External validation:** Evaluation on data separate from the development environment, ideally from different sites and conditions.
- **Prospective evaluation:** Assessment using data and workflows collected forward in time under real or planned use conditions.
- **Data drift:** Change in the distribution of model inputs or related observed data over time.
- **Score-distribution shift:** Change in the distribution of model outputs, probabilities, or risk scores.
- **Clinical performance degradation:** Decline in patient-relevant metrics such as sensitivity, specificity, calibration, false-positive rate, or subgroup parity.
- **Predetermined Change Control Plan:** FDA mechanism by which a manufacturer prespecifies certain future software modifications and their control methods.
- **FDA Sentinel:** Distributed surveillance system for monitoring the safety of regulated medical products; proposed here as a model for clinical AI oversight.
- **Federated evaluation network:** Cross-institutional monitoring in which sites perform standardized local evaluation and share coordinated results, potentially without centralizing patient-level data.
- **Coverage with evidence development:** CMS coverage approach that links payment to participation in evidence generation.
- **ARPA-H:** Advanced Research Projects Agency for Health.
- **Model provenance:** Documentation of training data, development process, versions, intended use, and known limitations.
- **Model retirement:** Controlled withdrawal of a model when it is obsolete, unsafe, unsupported, or no longer beneficial.

Questions and Verification

Open questions

- What minimum data volume is needed to detect degradation for rare outcomes or small demographic subgroups?
- Who owns the monitoring obligation when the vendor controls the model but the hospital controls the deployment data?
- How should surveillance handle proprietary models when meaningful evaluation requires access to inputs, versions, or training provenance?
- What is an acceptable delay between a prediction and an outcome label?
- How should thresholds balance patient safety with the risk of unnecessarily withdrawing a useful tool?
- Can the proposed network support non-device clinical algorithms that fall outside direct FDA oversight?

Claims to independently verify

- Current count of FDA-authorized AI/ML-enabled medical devices.
- The reported 43% share of recalls occurring in the first year after clearance.

- The reported 8.1% prospective-evaluation rate.
- The assertion in Figure 1 that 97% of devices use the 510(k) pathway.
- The demographic and transparency reporting rates.
- The retinal model's 5.5% to 46.6% scanner-related error increase.
- The precise extent of current FDA authority to mandate multisite validation and continuous monitoring.

Counterarguments and competing approaches

- Stronger premarket prospective trials could prevent some failures before deployment, though they cannot represent every future environment.
- Local health-system governance might be faster and more context-sensitive than a national network, but capacity would be uneven.
- Manufacturer-run monitoring may be technically efficient, but independent evaluation reduces conflicts of interest.
- A risk-tiered approach may be more feasible than identical continuous monitoring for every AI-enabled device.
- Excessive surveillance obligations could slow beneficial innovation or disadvantage smaller developers unless infrastructure and costs are shared.

Missing evidence or context

- Comparative cost estimates for the proposed network.
- A validated technical architecture and common data standard.
- Examples showing how often label-free drift alerts predict patient-relevant harm.
- Governance for cross-site signal adjudication and appeals.
- Quantified effects of monitoring on clinical outcomes, recall timing, or inequity.
- Operational detail for clinician notifications that informs without producing alert fatigue.

Logical next sources

1. The FDA's January 2025 lifecycle-management guidance for AI-enabled device software functions.
2. The FDA AI/ML-Based Software as a Medical Device Action Plan.
3. Technical and governance documentation for FDA Sentinel.
4. The cited studies on external validation, early recalls, demographic reporting, and device-summary transparency.
5. Research on data drift remedies, model calibration monitoring, and subgroup sequential testing.
6. Legal analysis of liability and foreseeability in clinical AI deployment.