NDSU PUBLIC HEALTH

October 14, 2025

Comment for Docket No. FDA-2025-N-4203

Measuring and Evaluating AI-Enabled Medical Device Performance in the Real World

Submitted by: Dr. Akshaya S. Bhagavathula, PharmD, PhD
Position: Associate Professor of Epidemiology, North Dakota State University
Expertise: Digital Epidemiology, Real-World Evidence, and AI-Driven Surveillance

Medical devices powered by AI aren't static tools - they're learning systems that continuously evolve as they encounter new data, patient populations, and clinical environments. From where I stand as an epidemiologist, we need to think about evaluating these devices the same way we monitor disease patterns in populations: dynamically and continuously, not as one-time assessments.

The fundamental question we're grappling with is whether these devices maintain their performance as circumstances change. Will a device that works well today still work well six months from now when patient demographics shift or clinical practices evolve? That's what I mean by temporal validity, and it's central to keeping patients safe.

1. Performance Metrics and Indicators

1a. What metrics or performance indicators are used to measure safety, effectiveness, and reliability of AI-enabled medical devices in real-world clinical use?

When we evaluate these devices in actual clinical settings, we need metrics that go well beyond simple accuracy or AUC scores. Real-world performance assessment requires metrics that account for population diversity, changes over time, and the uncertainty inherent in any prediction.

Here's what I believe should be standard practice:

- **Temporal stability tracking** – We need to monitor whether a device's calibration and discrimination capabilities hold steady over time or start to drift. This matters because what works well in January might perform differently by December as patient populations and clinical practices shift.
- **Fairness metrics across subgroups** – I recommend calculating performance differentials (changes in precision and recall) across demographic categories including sex, race and ethnicity, and social vulnerability index quintiles. Simply reporting overall performance masks important disparities.

- **Population-weighted scoring** – Using something like a population-weighted Brier score ensures that errors affecting underrepresented patient groups aren't simply averaged away in the overall statistics. Every patient population should count.

- **Clinical impact assessment** – We should weight different types of errors based on their actual clinical consequences. A false negative in cancer detection, for example, might warrant a weight five times higher than a false positive because the harm is so much greater.

- **Uncertainty reporting** – Every metric should come with confidence intervals and measures of model certainty expressed in probabilistic terms. Clinicians deserve to know not just what the algorithm predicts, but how confident we should be in that prediction.

All of these should be integrated into *post-market learning dashboards* that provide continuous monitoring rather than periodic snapshots.

1b. How are these metrics defined and weighted?

This is my core area of expertise. In my opinion, the weighting structure needs to reflect both clinical consequences and epidemiologic reality. Take diagnostic devices: depending on the condition, false negatives might need to be weighted much more heavily than false positives—sometimes by a factor of five or more.

I believe these weighting matrices should be developed collaboratively, bringing together data scientists who understand the statistical implications, clinicians who understand patient care consequences, and human factors specialists who understand how decisions actually get made in clinical settings. Whatever weighting scheme gets adopted should be transparently documented in the device's model card with clear justification.

1c. What timeframe defines "real-world clinical use" performance?

This is where I think we need to fundamentally shift our thinking. Monitoring shouldn't be time-bound – it should be continuous and adaptive, similar to how pharmacovigilance works for medications.

My recommendation is a phased approach:

- During the **first 12 months** (early-use intensive surveillance), conduct weekly drift analyses. This is when you're most likely to catch problems as the device encounters real-world variability.

- From **years 1-3** (stabilization phase), shift to quarterly calibration audits. By this point, you have a better sense of the device's baseline behavior, but you still need frequent checks.

- In the **mature phase** (year 3 onward), annual epidemiologic revalidation against evolving population data makes sense. But even here, continuous automated monitoring should be happening in the background.

The key point: *performance evaluation should never end*. These are learning systems operating in dynamic environments.

2. Real-World Evaluation Methods and Infrastructure

2a. What tools or processes are used to monitor post-deployment performance?

I envision a ***multi-layered approach*** that borrows heavily from pharmacovigilance models but adapts them for algorithmic systems:

Automated drift detection – Statistical process control charts can track calibration quality and sensitivity over time. When these metrics start decaying, you get an early warning signal.

Federated data infrastructure – By linking electronic health records, claims databases, and patient registries through federated approaches, we can monitor performance across diverse settings without requiring institutions to share sensitive raw data. Each site analyzes locally; only aggregated results flow back.

Regular explainability audits – Periodically testing the model using methods like SHAP analysis or counterfactual testing helps ensure the algorithm's reasoning hasn't fundamentally changed in unexpected ways.

Human factors observation – We should be tracking things like cognitive task load and workflow deviations. How is the device actually affecting clinical work patterns? Are clinicians getting overwhelmed or are they ignoring the system?

2b. How to balance human review vs automated monitoring?

I see these as complementary rather than competing approaches. Automation excels at identifying anomalies and flagging potential issues quickly. But interpreting whether those anomalies actually matter for patient safety requires human judgment.

The optimal design mirrors what we do in vaccine safety monitoring: *automated systems trigger alerts*, then expert panels made up of clinicians and data scientists review those triggers to

determine *causality and safety relevance*, followed by documented corrective actions. It's essentially adapting the VAERS model for algorithmic drift instead of adverse drug events.

2c. What infrastructure supports evaluation?

Several key components need to be in place:

1. **Federated learning environments** – Secure sandboxes where models can be retrained and tested under controlled conditions before any updates go live.
2. **Comprehensive audit logs** – Registries that capture input data distributions, model outputs, and actual user behavior. These create the paper trail necessary for meaningful retrospective analysis.
3. **Secure computational zones** – Using zero-trust architecture principles to enable reproducible audits while maintaining data security. This lets independent researchers verify findings without compromising privacy.
4. **Standardized metadata frameworks** – Everything should conform to established standards like HL7 FHIR and OMOP to ensure different systems can actually talk to each other.

3. Post-Market Data Sources and Quality Management

3a. What data sources are used?

The most complete picture comes from integrating what I call *real-world data fabrics*–weaving together electronic health records, device telemetry, insurance claims, patient-reported outcomes, and contextual digital data like mobility patterns or environmental exposures.

By linking these diverse sources, you can triangulate between what the device is doing and what's actually happening to patients. This kind of signal triangulation is essential for catching performance issues that might not be obvious from any single data stream.

3b. How to address data quality, completeness, and interoperability?

Data quality has to be transparent and measurable. I recommend requiring data lineage documentation that traces where information comes from and how it's been transformed. We should also establish quality indices with specific thresholds—for instance, requiring at least 85% completeness for key variables and ensuring data timeliness with less than 90-day lag.

On the technical side, probabilistic record linkage helps connect information across systems, bias-adjusted weighting addresses systematic gaps in the data, and mandatory conformity to standards like FHIR and DICOM-AI ensures different systems can actually exchange information meaningfully.

3c. Most effective methods for incorporating outcomes and user feedback?

Three strategies work particularly well in my experience:

- **Continuous outcome capture** – Rather than episodic surveys, link devices to registries that provide ongoing follow-up data about what happens to patients.
- **Adaptive learning loops** – Build mechanisms where clinician feedback gets incorporated directly into model retraining cycles through federated updates. This creates a genuine feedback loop rather than one-way communication.
- **Regular usability testing** – Every time you retrain the model, conduct mixed-methods usability studies to quantify whether the human-algorithm relationship is still working well. As the model evolves, the user experience needs to evolve appropriately too

4. Monitoring Triggers and Response Protocols

4a. What triggers additional assessment?

Several warning signs should prompt deeper investigation:

- **Performance degradation** – Any drop of 10% or more in calibration or sensitivity metrics warrants immediate attention.
- **Population shift** – When statistical tests show the input population has changed significantly (I'd use a Kolmogorov-Smirnov statistic threshold of 0.1 or higher), it suggests the model may be operating outside its validated range.
- **Outcome clustering** – If you see a surge in adverse clinical outcomes that coincides temporally with device outputs, that's a red flag demanding investigation.
- **Behavioral signals** – Extreme patterns in how clinicians use the device–either over-reliance or consistent under-utilization–suggest something may be wrong with the human-AI interaction.

4b. How to define and respond to degradation?

Statistical ***control limits*** provide clear definitions–Shewhart charts or CUSUM methods work well here. When limits are breached, the response protocol should be systematic:

Start with root cause analysis to understand what's driving the degradation. Then move to temporary model suspension while you investigate–patient safety comes first. Next, validate any fixes in shadow mode where the updated model runs alongside the current system without affecting patient care, letting you verify the problem is actually solved. Finally, notify regulators within 30 days of any significant performance issue.

5. Human–AI Interaction and User Experience

5a. How do usage patterns influence performance?

Here's something I've observed repeatedly: behavioral adaptation by users often matters more than algorithmic drift. The way clinicians interact with these systems evolves over time, and those changes can dramatically affect outcomes.

Over-trust leads to automation bias–clinicians stop thinking critically and just follow the algorithm's recommendations. Under-trust leads to inefficiency and wasted potential as clinicians ignore useful guidance. Both extremes are problematic.

That's why I advocate for longitudinally monitoring what I call trust *calibration curves*–tracking whether clinicians are developing appropriate levels of trust over time. Pairing this with ***cognitive load analytics*** (using tools like NASA TLX scores) and correlating those with error rates gives you a complete picture of the human-AI relationship.

5b. What design features and training strategies are most effective?

Several approaches show real promise:

- **Adaptive explainability** – The system should tailor its explanations based on who's using it. A cardiologist needs different kinds of rationale than a primary care physician. Context-aware explanation systems that match user expertise improve both trust and appropriate use.
- **Dynamic alert thresholds** – One major problem with AI systems is alert fatigue. Using contextual risk scoring to adjust when and how strongly the system alerts can prevent clinicians from tuning out.

- **Simulation-based training** – Creating digital twins that replicate adverse event scenarios lets clinicians rehearse their response in safe environments. This builds both competence and confidence.

- **Explainability literacy programs** – We need to actively teach clinicians about the interpretation limits of AI outputs. Understanding what the algorithm can and cannot tell you is foundational to appropriate use.

6. Additional Considerations and Best Practices

6a. Other considerations or best practices?

Several additional elements would strengthen the oversight framework:

1. **Integration with Quality Management Systems** – AI lifecycle monitoring shouldn't be separate from existing QMS processes. It should be embedded within them.

2. **Model card requirements** – Mandate transparent documentation covering training data epochs, feature sets, validation populations, and known biases. This should be as standard as drug labeling.

3. **Public equity dashboards** – Make subgroup performance data publicly available so disparities can't be hidden. Transparency drives accountability.

4. **FDA-coordinated learning networks** – Create structured ways for organizations to share anonymized information about drift patterns and bias findings. Individual organizations working in isolation won't catch systemic problems.

6b. Implementation barriers, incentives, privacy approaches

Let me be frank about the challenges we face when using AI:

- **Barriers** – Data fragmentation across healthcare systems, inconsistent liability frameworks that make organizations nervous about transparency, and lack of sustained funding for ongoing monitoring infrastructure.

- **Incentives** – The FDA could accelerate adoption through regulatory fast-track pathways for devices that implement robust continuous monitoring. Recognition within the Total Product Lifecycle program could also drive uptake.

- **Privacy** – We have the technical tools to do this right. Differential privacy mechanisms add mathematical noise that protects individuals while preserving population-level

insights. Secure multiparty computation within federated architectures lets institutions collaborate without exposing sensitive data. These approaches align with both HIPAA requirements and emerging AI governance frameworks.

AI-enabled medical devices aren't traditional devices–they're adaptive systems that require ongoing epidemiologic surveillance. Their safety depends on continuous monitoring, maintaining interpretability as they evolve, and ensuring equitable performance across all patient populations.

By integrating principles from **pharmacoepidemiology** (how we monitor drug safety), **digital health informatics** (how we manage and interpret health data), and **quality systems engineering** (how we build reliable processes), we can ensure AI in medicine develops transparently, reproducibly, and responsibly.

This isn't just about regulating technology. It's about building the infrastructure and practices that let us harness AI's potential while protecting patients and maintaining public trust.

I appreciate the opportunity to contribute to this important discussion and thank the FDA for soliciting input from the research and clinical communities on these critical issues affecting the future of healthcare.

I would be pleased to provide additional technical details or participate in expert discussions regarding these recommendations.

Please feel free to contact me at: akshaya.bhagavathula@ndsu.edu

Sincerely,

Dr. Akshaya Bhagavathula, PhD
Associate Professor of Epidemiology,
Department of Public Health | North Dakota State University
640S NDSU Dept 2662 | Fargo, ND 58108-6050
Email: Akshaya.bhagavathula@ndsu.edu; 701-231-6549