AMERICAN PSYCHIATRIC ASSOCIATION

800 Maine Avenue, S.W.
Suite 900
Washington, D.C. 20024

November 26, 2025

Docket No. FDA-2025-N-4203
"Artificial Intelligence (AI)–Enabled Medical Devices: Measuring Performance in Real-World Settings"

The American Psychiatric Association (APA) appreciates the opportunity to comment on the FDA's Request for Information (RFI) regarding AI-enabled medical devices and their evaluation in real-world clinical contexts. As the leading organization representing psychiatrists and advancing mental health innovation, APA strongly supports FDA's commitment to developing robust, transparent, and equitable frameworks for AI oversight.

Artificial intelligence–enabled medical devices are rapidly transforming clinical care, yet they remain governed by a fragmented and often inconsistent set of oversight mechanisms, if any at all. Many such devices reach patients without robust validation, post-market monitoring, or transparency around how they function overtime in diverse populations. In effect, the current regulatory landscape resembles a "digital frontier" in which innovation has outpaced accountability. The APA urges the FDA to bring greater structure, standards, and clinical rigor to this space and to establish clear protections for safety, privacy, effectiveness, and equity before and after market entry.

Below are APA's responses to the FDA RFI:

<u>Question 1: Metrics for Evaluating Real-World Performance</u>

RFI Text:
"What metrics or performance indicators do you use to measure the safety, effectiveness, and reliability of AI-enabled medical devices in real-world clinical use? How are these metrics defined and weighted when assessing different dimensions of performance and safety? What timeframe do you consider when evaluating real-world clinical use performance?"

APA recommends the adoption of a framework of standardized benchmarks for evaluating AI device performance that includes:

- Safety: Track adverse event rates, false-negative rates for critical conditions, and performance differences across demographic groups.
- Clinical Effectiveness: Measure sensitivity, specificity, predictive values, calibration, and clinical outcomes (e.g., time to diagnosis, symptom improvement).

- Ethical Responsibilities: Informed consent, mitigating bias, secure and private data use, transparent models and use across the service.
- Performance Reliability for medical use: Monitor consistency over time and across clinical sites, including system uptime and clinician usability indicators such as override rates or alert fatigue.

APA notes that many generative-AI devices will not be used as medical devices and therefore will fall out of the purview of the FDA's enforcement. However, these devices may be marketed as wellness devices or could be mimicking treatment options for patients with symptoms of mental health conditions. These devices, if there is public use, should be evaluated and monitored with a responsible standard set for measuring harms such as suicidal ideation.

Weighting and timeframe: Metrics should be weighted by device risk class (from low to unacceptable) prioritizing those with highest weight for high-risk applications. Evaluations should begin within 3 months and then continuously with updates and changes to the foundational model.

Question 2: Data Sources for Real-World Performance Evaluation

RFI Text:
"What data sources and evidence types are most useful for evaluating real-world performance of AI-enabled devices? How can FDA encourage access to or use of such data sources?"

APA supports FDA's use of diverse, high-quality data sources to capture AI performance in varied clinical environments. These may include:

- Electronic Health Records (EHRs) and device logs to track usage patterns and outcomes;
- Demographic and epidemiological data;
- Deidentified conversation transcripts and video/voice recordings;
- Patient-reported outcomes to assess quality of life and care satisfaction;
- Clinician feedback systems to document usability issues, outcomes, and unintended consequences;
- Synthetic data used when real-world data is too sensitive or needs extra privacy protections; and
- Post-market surveillance databases to aggregate de-identified safety data across sites.

To promote access, FDA should enforce mandatory data-sharing frameworks and interoperability standards that protect patient privacy while enabling oversight.

Question 3: Monitoring, Updating, and Managing Model Drift

RFI Text:
"How should performance of AI-enabled devices be monitored, updated, or managed over time to ensure continued safety and effectiveness in real-world use?"

Developers and health systems should maintain clear, ongoing procedures for checking whether AI tools continue to function safely and effectively after implementation.

APA recommends:

- Regular review of performance: AI systems must be monitored continuously with alert systems in place to notify health systems of possible drift or changes in safety or accuracy. Testing frameworks and reviews must evolve with the models to ensure that any drift is accounted for overtime. A data dashboard or other monitoring system must be reviewed at regular intervals throughout the year for maintenance and performance checks.
- Early warning indicators: Establish clear triggers for review when an AI system's behavior changes in ways that could compromise safety or accuracy. This can include examples such as a sudden drop in performance metrics, inconsistent or implausible recommendations, or repeated user reports of confusing outputs. Systems should also include an **"**emergency stop" or suspension mechanism that allows temporary deactivation of the AI function if concerning patterns are detected until the issue is investigated and resolved.
- Update and approval procedures: When models are modified, developers should document what changed, why, and how it was validated. FDA should clarify when such updates require additional review.
- Human oversight: Clinicians must be able to question or override AI outputs and should be notified of significant updates.
- Transparent reporting: Share summaries of ongoing performance monitoring and updates with FDA and the public, without identifying patient data.

These steps ensure AI devices remain safe, clinically valid, and trustworthy throughout their lifecycle.

<u>Question 4: Transparency, Explainability, and Human Oversight</u>
RFI Text:
"What approaches promote transparency, interpretability, and clinician understanding of AI-enabled device recommendations during real-world use?"

Transparency is critical to clinical safety and trust. APA recommends that developers and regulators ensure:

- Clear documentation of data sources utilized in addition to the foundational model, model design, and intended use;

        - Future development of a stakeholder consensus framework supported by the FDA
    - Plain-language explanations of outputs (e.g., why a result was generated, how confident the model is); and
    - Guidelines for clinician oversight, including when and how to override AI recommendations.
    - A real-time adverse reporting system that allows post-market review of AI-model drift and review of adverse events of end-users when known

AI systems should support not replace clinical judgement of a psychiatrist. This is especially vital in psychiatric practice, where care decisions require empathy, context, and individualized judgment.

Question 5: Equity and Subpopulation Performance

RFI Text:
"How can real-world performance assessments of AI-enabled devices incorporate considerations of equity and performance across subpopulations?"

APA recommends that FDA require regular analysis of performance across demographic and clinical subgroups to identify inequities. Key practices include:

- Evaluating error rates by race, ethnicity, gender, age, and comorbidity;
- Ensuring training and validation datasets reflect the populations served; and
- Reporting any disparities and remediation steps in post-market updates.

In psychiatry, data bias can amplify social inequities such as increased exposure to environments where substance use is common, trauma-related disorders, increased risk of common mental disorders, and increased rates of suicidal thoughts, attempts, and self-harm; therefore, equity metrics should be treated as a safety requirement, not an optional feature.

Question 6: Collaboration and Standardization

RFI Text:
"What collaborative approaches or data standards could improve FDA's ability to evaluate and monitor AI-enabled device performance in real-world settings?"

APA supports collaboration between FDA, professional medical societies, academic institutions, and patient groups to standardize performance measurement. Recommendations include:

- Establishing shared templates for performance reporting and bias evaluation;

- Creating multidisciplinary advisory panels that include psychiatrists, data scientists, and ethicists; and
- Aligning with federal AI initiatives promoting interoperability, transparency, accountability, and fairness.

The APA commends FDA for its leadership in advancing responsible AI oversight. We urge adoption of a risk-based, lifecycle-oriented, and equity-focused framework to ensure AI-enabled medical devices are safe, effective, and trusted by clinicians and patients alike. APA stands ready to collaborate with FDA to develop psychiatric-specific guidance on AI evaluation and real-world performance monitoring. Please contact Zuhal Haidari (zhaidari@psych.org) with any questions or for more information.

Sincerely,

Marketa Wills, MD, MBA, FAPA
CEO and Medical Director
American Psychiatric Association