December 1, 2025

The Honorable Martin A. Makary, MD
Commissioner
Food and Drug Administration
10903 New Hampshire Avenue
Silver Spring, MD 20993

Submitted electronically via regulations.gov

RE: Docket No. FDA-2025-N-4203; Measuring and Evaluating Artificial Intelligence-enabled Medical Device Performance in the Real World; Request for Public Comment

Dear Commissioner Makary:

On behalf of the American Academy of Family Physicians (AAFP), representing 128,300 family physicians and medical students across the country, I write in response to the Food and Drug Administration's (FDA) request for public comment on measuring and evaluating artificial intelligence (AI)-enabled medical device performance in the real-world. The AAFP strongly believes in the importance of real-world data (RWD) and testing, and we appreciate the agency seeking input on how to best measure and evaluate the performance of AI-enabled medical devices through RWD and feedback. The AAFP is committed to ensuring that AI technologies and medical devices are integrated into health care in ways that are safe, effective, fair, and transparent.

The family medicine experience is based on a deeply personal patient-physician interaction that requires support from technology, including artificial intelligence. Therefore, in 2023, the AAFP developed an initial set of principles that we believe must be followed for AI technologies if they are to be applied to family medicine. The AAFP believes AI tools should be evaluated with the same rigor as any other tool utilized in health care, and we believe AI has the potential to support the four C's of primary care: first contact, comprehensiveness, continuity, and coordination of care.

Similarly, the AAFP is extremely supportive of real-world testing and the use of RWD in evaluating new technologies, and we believe it is critical to ensure all health information technology (IT) products and tools facilitate their intended uses without negative or unintended consequences. Health care is a complex, adaptive system that cannot always be predicted, which means real-world testing must be conducted, and RWD must be considered. We appreciate FDA working with the Assistant Secretary for Technology Policy/Office of the National Coordinator for Health Information Technology (ASTP/ONC) to align health IT policies throughout the Department of Health and Human Services (HHS), including this discussion of how

1133 Connecticut Ave., NW, Ste. 1100
Washington, DC 20036-1011

info@aafp.org
(800) 794-7481
(202) 232-9033

www.aafp.org

RWD could be utilized to detect, assess, and mitigate AI-enabled medical devices' performance changes over time.

Question 4a: Monitoring Triggers and Response Protocols: What triggers the need for additional assessments and more intensive evaluation?

The AAFP recommends expanding upon the foundational 2014 work *Health IT and Clinical Decision Support Systems: Human Factors and Successful Adoption*, which outlined early principles for monitoring clinical decision support systems.[i] This journal article provides the bedrock for a risk-based framework for software in health care, and it remains valid with the explosion of AI tools in health care. The work still needed by the health IT community is to establish boundaries within the risk framework to classify an AI tool or solution as low-risk (e.g., green), moderate-risk (e.g., yellow), high-risk (e.g., red), or extreme-risk (e.g., black).

Additionally, here are some specific triggers we recommend FDA consider:

1. **Performance and outcome drift**: Statistically significant degradation in prospective performance (e.g., accuracy, calibration, positive predictive value, and negative predictive value (PPV/NPV)) or worsening patient outcomes, as compared to pre-deployment baselines or pre-specified control limits.[ii]
2. **Population and practice drift**: Meaningful changes in case-mix, geography, payer mix, devices, or workflows, especially if subgroup performance diverges.
3. **Adoption signals** (for clinical decision support and large language model (LLM) tools used in workflow): When end-users or organizations fail to adopt AI tools or solutions, this should trigger a re-evaluation of the solution's real-world fit.
4. **High or rising alert override rates**: Alert overload is a safety issue; a trigger should be set for increased alert and override rates beyond established thresholds, including sudden spikes in interruptive alert volume ("over-triggering") or time-to-decision support (latency) increases, both of which are associated with alert fatigue and safety risks.[iii]
5. **Safety, reliability, or usability signals**: Adverse events and near misses, clinician hazard reports, patient complaints, reproducibility failures, user interface (UI) regressions, unavailability, or hallucination-like outputs for LLM-based tools.
6. **All model updates**: Such as a change in weights of factors, features, data pipeline changes, an EHR upgrade, or a vendor swap taking place without prior real-world re-verification and user re-training.
7. **External knowledge changes**: New evidence, guidelines, or drug approvals that invalidate AI assumptions (e.g., contraindications, new dose ranges).

Question 5b: Human-AI Interaction and User Experience: What design features, user training, or communication strategies have proven most effective for maintaining safe and effective use as systems evolve?

The AAFP strongly believes that physician input must be embedded throughout the AI lifecycle—from discovery and design to deployment and post-market surveillance. Effective strategies include:

- Co-design processes that involve frontline clinicians to ensure usability and workflow integration;
- User training programs tailored to the clinical context, emphasizing limitations and appropriate use; and
- Transparent communication strategies that provide real-time explanations of AI recommendations and confidence levels.

Question 6b: Additional Considerations and Best Practices: Please address any implementation barriers encountered, incentives that supported your efforts, and approaches to maintaining patient privacy and data protections.

The AAFP urges FDA to review the AAFP and Rock Health's recent report The Starfield Signal: A Shared Vision and Roadmap for AI in Primary Care and the JAMA Special Communication for comprehensive guidance on implementation barriers and best practices.[iv] Key considerations include:

- Implementation barriers such as misaligned payment models, fragmented infrastructure, and trust deficits. Best practices include leveraging direct primary care and value-based payment for addressing misaligned payment models, in addition to new payments to support AI deployment and use.
- Starting with well-designed pilots that have clearly defined metrics and thresholds to determine whether to proceed with a full-scale rollout.
- Governance frameworks and committees within health care organizations, as well as rigorous certification programs to encourage responsible adoption.
- Privacy and data protection: enforce robust de-identification standards, consent processes, and regulatory compliance with the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and emerging state regulations.

In conclusion, the AAFP reiterates its strong support for both the responsible adoption of AI tools in health care and the use of RWD to measure and evaluate those tools. Thank you for the opportunity to provide written comments on this important topic and recommendations on how these tools can be appropriately utilized and safeguarded. We appreciate FDA's leadership in shaping the future of AI in health care and stand ready to collaborate on developing policies and frameworks

that ensure safe and effective AI integration and evaluation through RWD and feedback. For more information or questions, please contact Mandi Neff, Senior Strategist, Regulatory and Policy, at mneff2@aafp.org.

Sincerely,

Jen Brull, MD, FAAFP
Board Chair
American Academy of Family Physicians

---

[i] L Ohno-Machado, Health IT and clinical decision support systems: human factors and successful adoption, *Journal of the American Medical Informatics Association*, Volume 21, Issue e2, October 2014, Page e180, https://doi.org/10.1136/amiajnl-2014-003279.

[ii] Angus DC, Khera R, Lieu T, et al. AI, Health, and Health Care Today and Tomorrow: The JAMA Summit Report on Artificial Intelligence. JAMA. 2025;334(18):1650–1664. doi:10.1001/jama.2025.18490.

[iii] Karen C Nanji, Diane L Seger, Sarah P Slight, Mary G Amato, Patrick E Beeler, Qoua L Her, Olivia Dalleur, Tewodros Eguale, Adrian Wong, Elizabeth R Silvers, Michael Swerdloff, Salman T Hussain, Nivethietha Maniam, Julie M Fiskio, Patricia C Dykes, David W Bates, Medication-related clinical decision support alert overrides in inpatients, *Journal of the American Medical Informatics Association*, Volume 25, Issue 5, May 2018, Pages 476–481, https://doi.org/10.1093/jamia/ocx115.

[iv] Angus DC, Khera R, Lieu T, et al. AI, Health, and Health Care Today and Tomorrow: The JAMA Summit Report on Artificial Intelligence. JAMA. 2025;334(18):1650–1664. doi:10.1001/jama.2025.18490.