



Submitted Electronically via Regulations.gov

October 19, 2026

Michelle Tarver, MD, PhD
Director
Center for Devices and Radiological Health
U.S. Food and Drug Administration
10903 New Hampshire Avenue
Silver Spring, MD 20993

Re: Docket No. FDA-2026-N-7874—Considerations for the Regulation of Generative AI-Enabled Medical Devices: Discussion Paper and Request for Feedback

Dear Dr. Tarver:

The American Medical Informatics Association (AMIA) appreciates the opportunity to provide feedback to the U.S. Food and Drug Administration's Center for Devices and Radiological Health (CDRH) on its discussion paper, Considerations for the Regulation of Generative AI-Enabled Medical Devices. AMIA is the professional home for more than 6,400 clinicians, researchers, educators, data scientists, and health technology leaders dedicated to advancing health and healthcare through informatics. Our members develop, evaluate, implement, and oversee health information systems and artificial intelligence-enabled technologies across clinical care, research, public health, and education. Drawing on this multidisciplinary expertise, AMIA supports a risk-proportionate, evidence-based, and total product lifecycle approach that protects patients, promotes transparency and accountability, enables interoperability and meaningful real-world evaluation, and supports responsible innovation.

Scope and Definition of GenAI Enabled Medical Devices

AMIA recommends that CDRH clarify which generative AI-enabled functions qualify as device functions and therefore fall within FDA's regulatory oversight. Existing guidance, most notably FDA's "Policy for Device Software Functions and Mobile Medical Applications"¹ and "Clinical Decision Support Software"² guidance, provides numerous examples of software functions that are and are not the focus of oversight. However, those examples, including the January CDS guidance, do not

¹ U.S. Food and Drug Administration. (2022, September 28). *Policy for device software functions and mobile medical applications: Guidance for industry and Food and Drug Administration staff*.

<https://www.fda.gov/media/80958/download>

² U.S. Food and Drug Administration. (2026, January 29). *Clinical decision support software: Guidance for industry and Food and Drug Administration staff*. <https://www.fda.gov/media/109618/download>

directly address generative AI-enabled functions, leaving industry stakeholders and developers to infer how the existing policies apply.

Many issues raised in this discussion paper are also relevant to widely adopted tools that AMIA believes are non-device software functions. AMIA's responses would likely differ for this broad group of non-device AI-enabled functions and for the narrower set of AI-enabled device software functions subject to FDA oversight. Concrete examples, parallel to those in existing guidance, would help stakeholders understand which technologies FDA intends the proposed frameworks to cover and provide more focused feedback.

Illustrative examples that we believe fall within scope, depending on their intended use and functionality, include:

- Generative imaging reconstruction that creates diagnostic-quality images from undersampled data;
- Multimodal deterioration-detection systems that analyze waveforms, laboratory results, and clinical notes and generate patient-specific narratives or recommended actions;
- Patient-facing applications that interpret symptoms or physiologic data and recommend a specific medication dose; and
- AI-enhanced patient monitoring devices that convert measurements into patient-specific diagnostic, treatment, or escalation recommendations.

Illustrative examples that we believe fall outside of scope include:

- Recording of a clinical encounter, AI-driven transcription, and drafting of a clinical note, potentially inclusive of drafting orders for physician review
- Summarization of a patient's chart so long as it is not inclusive of near real-time signal data.
- AI-driven evidence-retrieval tools, including Open Evidence and commercially available foundation models, whether used by clinicians or patients, so long as the outputs meet the general examples provided in the CDS guidance.

FDA should clarify when technologies such as these become subject to oversight. CDRH should map representative GenAI functions to the statutory definition of a device and to each criterion for non-device clinical decision support, including examples showing when patient-specific outputs, directives, clinical claims, or influence on diagnosis or treatment bring a function within FDA oversight. This clarification would ensure that stakeholders evaluate the same categories of technology and would provide the necessary foundation for the postmarket infrastructure recommendations that follow.

Section IV: Considerations for the Assessment of Risk for GenAI-Enabled Devices

Question 1: Does the two-axis risk framework, organized around device activity and the consequence of relying on an incorrect output, appropriately capture the dimensions most relevant to the risk of a GenAI-enabled software function? If there are additional dimensions—such as the reversibility of a resulting action, the availability of downstream safeguards, the time pressure of the deployment setting, or the traceability of the output (i.e., to primary source materials)—that should be represented in a risk framework, please describe and provide examples of how they should be represented.

AMIA agrees that device activity and the consequences of an incorrect output are relevant, but the proposed axes should not be treated as a complete risk model. CDRH should anchor classification in the probability and severity of harm, then apply a limited set of contextual modifiers that can raise or lower either dimension. This approach keeps the framework usable without reducing a multidimensional problem to categorical labels such as non-directive, directive, or action-taking.

- The probability that an incorrect or misleading output will be relied upon before correction;
- The reversibility of harm if the output is relied upon;
- Whether the output can be independently reviewed, challenged, and overridden;
- The intended user's relevant expertise, health literacy, and ability to recognize an error;
- Potential influence on the user, highlighted by the strength of the recommendation, degree of personalization, and cumulative effect of multi-turn interaction;
- The autonomy of the system and its operation;
- The quality and traceability of supporting evidence, including evidence grade where available, source fidelity, and uncertainty communication; and the
- The scale of deployment and potential population impact.

FDA should use these dimensions to establish transparent risk tiers and illustrative examples. The framework should not presume that informational functions are inherently safer than action-directing functions since contextual information provision itself can be persuasive or that professional expertise eliminates automation bias. CDRH also should not assume that risk rises linearly or exponentially with autonomy. Autonomy can create nonlinear risk through scale, compounding errors, and fewer opportunities to intervene, but some autonomous functions may reduce human error. Classification should therefore rely on empirical performance in the intended workflow and under expected fallback conditions.

Human review should count as a risk-reducing safeguard only when the reviewer has the time, information, clinical and AI expertise, and authority to act, and when human and model failure modes are not so correlated that review merely confirms the same error. Where possible, evaluation should compare the device alone, the unaided clinician, and the human-AI team.

Below is an example of a multidimensional risk framework with 4 stages to determine a device's risk.

1. Assess the probability and severity of harm

First, assess the device using two core dimensions that align with established risk-assessment practice:

- Probability of harm: How likely is the device to generate an incorrect, misleading, or incomplete output?
- Severity of harm: If an incorrect output is relied upon, could it cause inconvenience, unnecessary care, delayed treatment, serious injury, or a life-threatening outcome?

Baseline risk increases as both the likelihood of harm and the severity of that harm increase; the relationship should not be assumed to be linear. As a contextual modifier, the risk increases if there is high probability of reliance before correction.

For example:

- A tool that summarizes general clinical information may begin in a lower-risk area only when errors are unlikely to cause harm, any potential harm is not severe, and users can readily verify the source information with minimal additional effort.
- A patient-facing tool that recommends whether to seek emergency care would begin in a higher-risk area because an incorrect output, even if improbable, is likely to be relied upon without correction. Additionally, it may prompt harmful outcomes: either delayed care or unnecessary emergency care utilization.
- A system that recommends or autonomously changes an insulin dose would begin in the highest-risk area because an error could produce immediate and severe harm.

This placement establishes the baseline risk, not the final classification.

2. Apply AMIA's contextual risk modifiers

Next, assess the circumstances in which the device will operate. The same contextual modifiers identified above should be applied consistently, with each modifier potentially increasing or decreasing the estimated probability or severity of harm:

- **Reversibility of harm if the output is relied upon:** Can an action or resulting harm be detected and reversed before serious or lasting consequences occur?
- **Ability to independently review, challenge, and override:** Can a qualified person meaningfully verify, challenge, stop, or override the device before harm occurs, and does empirical testing demonstrate that human review corrects errors rather than amplifying shared reasoning flaws or biases?
- **Intended user's relevant expertise, health literacy, and ability to recognize an error:** Is the output delivered to a patient, caregiver, generalist, specialist, or care team, and has the

device been evaluated for the intended users across relevant languages, literacy levels, and clinical settings?

- **Potential influence on the user:** How strongly does the output recommend an action, how personalized is it, and could conversational, anthropomorphic, or multi-turn interactions increase reliance over time?
- **Autonomy of the system and its operation:** Does the device provide information, recommend an action, initiate an action, or autonomously execute a multi-step task, and where are meaningful opportunities for human intervention?
- **Urgency of the decision and availability of downstream safeguards:** Is the decision time-sensitive, and are independent safeguards available to identify and correct an error before harm occurs?
- **Quality and traceability of supporting evidence:** Can users assess the evidence supporting the output, its source fidelity, applicable evidence grade, limitations, and validated uncertainty information, and is an appropriate audit trail available?
- **Scale of deployment and potential population impact:** Could widespread use amplify errors, inequities, workflow disruption, automation bias, privacy risks, or cybersecurity vulnerabilities, particularly across historically underserved populations or varied clinical and technological environments?

These modifiers help prevent two devices with similar intended functions from receiving the same risk classification when their operating conditions and potential effects differ materially.

As a secondary aid for applying two of these same modifiers, CDRH could use a simple matrix showing the relationship between an output's potential influence on the user, particularly the strength of its recommendation, and the quality and traceability of its supporting evidence. A strong, personalized recommendation based on weak or unverifiable evidence should increase the assessed risk even when the output is labeled informational. Conversely, a less directive output supported by strong, traceable evidence may present lower interaction risk. This matrix should help apply the contextual modifiers, not replace the underlying probability-and-severity assessment.

3. Assign an overall risk tier

After combining the baseline matrix placement with the contextual modifiers, the reviewer assigns an overall risk tier. The table below illustrates how the two core axes establish a preliminary risk level and how contextual modifiers may move a device higher or lower. An upward arrow (↑) indicates potential upward pressure on risk, while a downward arrow (↓) indicates potential risk mitigation. No single modifier should automatically determine the final tier.

Risk Tier	Probability and severity of harm	Contextual modifiers influencing placement	Illustrative example
-----------	----------------------------------	--	----------------------

Lower	Low probability and limited severity of harm.	↓ Reversibility: Effects are readily corrected. ↓ Independent review: Users can verify the output before relying on it. ↓ Influence: The output is non-directive. ↓ Evidence: Sources and uncertainty are clearly presented. ↓ Scale: Deployment is limited or controlled.	A clinician-facing tool summarizes general clinical information, cites its sources, does not recommend or initiate an action, and is reviewed before use. Subgroup testing identifies no material performance differences among intended users.
Moderate	Moderate probability or severity of harm, with generally manageable or reversible effects.	↑ Intended user: A patient, caregiver, or generalist may have difficulty recognizing an error. ↑ Influence: The output is personalized or recommends an action. ↓ Reversibility and safeguards: Effects are reversible, and clinician review or another safeguard is available. ↑ Equity: Usability or performance varies across language, literacy, disability, or demographic groups.	A patient-facing tool recommends routine clinical follow-up. A clinician can review the recommendation before treatment changes occur, but testing identifies reduced comprehension or accuracy for users with limited English proficiency.

High	Meaningful probability of serious harm or a lower-probability error with potentially serious consequences.	↑ Independent review: Verification is difficult or unlikely before reliance. ↑ Urgency and safeguards: The decision is time-sensitive, with few downstream checks. ↑ Influence: The output strongly directs an action. ↑ Evidence: Supporting evidence is incomplete or difficult to verify. ↑ Scale and equity: Broad deployment could amplify subgroup-specific errors or disparities.	A patient-facing emergency-triage tool provides strongly worded, personalized recommendations that users are likely to follow immediately. Testing identifies higher under-escalation rates for a clinically relevant subgroup, and deployment across multiple health systems could expose many patients to the disparity.
Very High	Plausible severe, irreversible, or life-threatening harm.	↑ Reversibility: Harm may be difficult or impossible to reverse. ↑ Autonomy: The system initiates or executes clinical actions. ↑ Urgency and safeguards: There is little opportunity to intervene before harm occurs. ↑ Compounding actions: Multi-step outputs can propagate an initial error. ↑ Scale and population impact: System-wide deployment could produce widespread harm or magnify inequitable outcomes.	An autonomous agent changes insulin or anticoagulant dosing and executes related orders across a health system without a mandatory human checkpoint. An error could cause immediate serious harm, propagate through multiple actions, and disproportionately affect a subgroup for which performance is lower.

Equity is a cross-cutting consideration within the intended-user and population-impact modifiers. Materially worse usability, performance, or outcomes for a clinically relevant subgroup may justify a higher risk tier even when average performance appears acceptable.

The framework does not necessarily require a numerical score. FDA could instead publish decision criteria and examples showing when particular modifiers should move a device into a higher or lower tier.

4. Match oversight to the risk tier

The assigned tier should determine the rigor of regulatory evidence, safeguards, and monitoring. However, usability testing, subgroup analysis, standardized AI exposure and provenance tracking, basic transparency, complaint and adverse-event monitoring, and periodic reassessment should apply across all tiers. These measures should help validate the initial risk classification and trigger reassignment when real-world evidence indicates greater or lesser risk. This is important because usability problems and unequal performance can increase the risk of otherwise beneficial technologies and contribute to intervention-generated inequalities.³

The same interoperable exposure and provenance infrastructure should support every tier, although the granularity of data collection, frequency of review, and intensity of regulatory oversight should increase with risk:

- **Lower-risk devices:** Streamlined validation proportionate to the intended use; baseline usability testing and subgroup analysis; disclosure of intended use, limitations, and known risks; and periodic review of exposure, performance, complaint, and safety data.
- **Moderate-risk devices:** Lower-risk measures conducted with greater rigor, supplemented by representative benchmarking, scenario-based usability testing, clinically relevant subgroup analyses, clinical confirmation, defined human-oversight procedures, and more frequent performance review.
- **High-risk devices:** Moderate-risk measures supplemented by prospective clinical evidence, independent evaluation, robust human-factors testing, continuous performance monitoring, detailed audit logs, strict change controls, and procedures for rapidly escalating identified safety or equity concerns.
- **Very-high-risk devices:** High-risk measures supplemented by especially strong premarket evidence, explicit authorization and operating limits, mandatory human checkpoints where feasible, rapid shutdown or rollback capabilities, intensive postmarket surveillance, and frequent reassessment of usability and performance across relevant populations.

Question 2: CDRH seeks input on how to account for the spectrum of GenAI-enabled informational functions that vary in the degree to which they direct a user to a particular action (i.e., between “non-directive” and “action-directing”). What characteristics of an output—such as its wording, specificity, personalization, or context—could be considered as modifiers of the risk associated with an informational function, after accounting for the device’s overall functionality and intended use? What additional information would provide manufacturers with sufficient clarity and predictability around risk assessment for

³ Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. *J Am Med Inform Assoc*. 2018;25(8):1080–1088. doi:[10.1093/jamia/ocy052](https://doi.org/10.1093/jamia/ocy052)

informational functions while recognizing that directiveness may exist along a continuum rather than as a binary distinction?

AMIA cautions against treating directiveness as a binary characteristic or a reliable stand-alone proxy for risk. The distinction between informational and action-directing output cannot be determined from wording or a sponsor's label alone. An output framed as informational may still cause action or inaction^{4,5} when it is personalized⁶, long or repeated⁷, presented with unwarranted certainty⁸, or delivered through a conversational interface that develops rapport and trust^{9,10}. Risk should reflect the probability and severity of harm in actual use and the output's foreseeable influence on a clinical decision.

FDA should provide examples addressing specificity, personalization, urgency, imperative language, recommendation strength, the quality and traceability of supporting evidence, validated uncertainty, anthropomorphic design, cumulative conversational context, and the likelihood that an output will alter care. A strong recommendation based on weak or unverifiable evidence should be treated as a meaningful risk modifier. Directiveness should be evaluated across the full interaction, and a disclaimer alone should not convert action-directing content into a low-risk informational function.

CDRH should use usability and human-factors evidence to test how intended users interpret and act on outputs under realistic conditions. This evidence is more reliable than attempting to classify a device solely from the formal wording of a single response.

Question 3: CDRH seeks input on whether and when a GenAI-enabled function that results in delivery of clinical information to patients, as opposed to HCPs, could present different or

⁴ Bickmore TW, Trinh H, Olafsson S, O'Leary TK, Asadi R, Rickles NM, Cruz R. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. *Journal of medical Internet research*. 2018 Sep 4;20(9):e11510. <https://pubmed.ncbi.nlm.nih.gov/30181110/>

⁵ Shekar, S., Pataranutaporn, P., Sarabu, C., Cecchi, G. A., & Maes, P. (2025). People Overtrust AI-Generated Medical Advice despite Low Accuracy. *NEJM AI*. <https://ai.nejm.org/doi/full/10.1056/Aloa2300015>

⁶ Salvi, F., Horta Ribeiro, M., Gallotti, R. et al. On the conversational persuasiveness of GPT-4. *Nat Hum Behav*, 1645–1653 (2025). <https://doi.org/10.1038/s41562-025-02194-6>

⁷ Steyvers M, Tejeda H, Kumar A, Belem C, Karny S, Hu X, Mayer LW, Smyth P. What large language models know and what people think they know. *Nature Machine Intelligence*. 2025 Feb;7(2):221-31. <https://doi.org/10.1038/s42256-024-00976-7>

⁸ Zhou K, Hwang J, Ren X, Sap M. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 2024 Aug (pp. 3623-3643). <https://aclanthology.org/2024.acl-long.198/>

⁹ Li Q, Luximon Y, Zhang J. The Influence of Anthropomorphic Cues on Patients' Perceived Anthropomorphism, Social Presence, Trust Building, and Acceptance of Health Care Conversational Agents: Within-Subject Web-Based Experiment. *J Med Internet Res* 2023; <https://www.jmir.org/2023/1/e44479>

¹⁰ Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies. W. Shi, X. Wang, Y. Oh, J. Zhang, S. Sahay, and Z. Yu. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3313831.3376843>

higher risks, while also recognizing the potential benefits associated with improved patient empowerment, engagement, and access to clinical information. What device characteristics, output features, or safeguards might mitigate risks that could arise when a user lacks the domain knowledge to independently evaluate an output, without unnecessarily underestimating patient capability?

Patient-facing functions should not automatically be classified as higher risk, but risk increases when delivered by a computer application imbued with authority based on who and where it is delivered¹¹, outputs are personalized, action-directing, difficult to verify, or likely to influence urgent clinical decisions. Human-like avatars, voice and conversational interfaces may also increase perceived authority and trust and should be considered in the risk assessment^{8,12}.

Safeguards should include plain-language disclosure that the user is interacting with AI, communication of uncertainty, citations or source provenance, clear scope limitations, emergency-escalation protocols, access to human assistance, privacy protections, and meaningful opt-in and opt-out choices. Patient-facing devices should be tested with equal rigor across supported languages, literacy levels, disabilities, and infrastructure conditions, including low-bandwidth environments. Patients and caregivers should participate in design and evaluation so that safeguards reflect real-world health and digital literacy needs.

Question 4: CDRH seeks input on whether and how the distinction between generalist and specialist physicians might inform the assessment of risk for HCP-facing GenAI-enabled functions. Under what circumstances, if any, might risk be affected when an HCP who lacks the relevant clinical specialist knowledge receives information that falls within a specialist area of practice? What device characteristics or safeguards might mitigate such risks?

The distinction between generalist and specialist users should inform, but not independently determine, risk. FDA should evaluate whether the intended user can identify an incorrect output, interpret uncertainty, and recognize when specialist consultation is required. Expertise alone should not be treated as a sufficient safeguard because specialists may also over-rely on AI outputs under time and workflow pressures.

Devices should be tested with the clinicians expected to use them, not only subject-matter experts. Safeguards could include source traceability, empirically calibrated uncertainty information, clear scope limitations, referral recommendations, and mandatory secondary review for high-consequence outputs or when validated thresholds are not met. For example, an AI-generated

¹¹ Kopka M, Schmieding ML, Rieger T, Roesler E, Balzer F, Feufel MA. Determinants of Laypersons' Trust in Medical Decision Aids: Randomized Controlled Trial. *JMIR Hum Factors* 2022;9(2):e35219 <https://humanfactors.jmir.org/2022/2/e35219>

¹² Cohn M, Pushkarna M, Olanubi GO, Moran JM, Padgett D, Mengesha Z, Heldreth C. Believing anthropomorphism: examining the role of anthropomorphic cues on trust in large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* 2024 May 11 (pp. 1-15). <https://dl.acm.org/doi/10.1145/3613905.3650818>

imaging interpretation should not replace independent review of the underlying image when the intended workflow requires professional confirmation.

Question 5: For multi-turn conversational GenAI-enabled devices that may migrate from providing “non-directive” information to “action-directing” information over the course of an exchange, how should risk be assessed across realistic conversational trajectories? How could the intended use of such a device be characterized when its behavior is emergent across a conversation?

FDA should evaluate the complete conversational trajectory, including reasonably foreseeable situations in which a device moves from general information to personalized or action-directing advice. The assessment should account for increasing rapport and trust over multiple turns and for errors that accumulate when later responses build on an earlier incorrect premise.

Testing should include branching scenarios, repeated questioning, incomplete or contradictory information, attempts to move the device outside its intended scope, and conversations involving urgent symptoms. Risk should reflect the highest-risk behavior reasonably expected across the interaction, rather than only the function performed at the first turn.

FDA should require a specific intended-use envelope and test whether the device consistently maintains those boundaries. Behavior described as emergent should not excuse recurring scope drift into unvalidated diagnosis, treatment, or other off-label functions; frequent boundary failures should be treated as a material safety concern.

Section V: A Competency-Based Approach for Premarket Evaluation of GenAI Enabled Devices

Question 7: Is the competency-based approach described above, i.e., device benchmarking followed by clinical confirmation, a useful and appropriate framework for evaluating GenAI-enabled devices?

AMIA supports competency-based evaluation as a useful component of FDA review, but benchmarking should complement, not replace, clinical validation.

FDA should evaluate the final configured product within its intended workflow, including its interaction with clinicians, patients, EHRs, external tools, and institutional safeguards. Evaluation should address clinical performance, health equity, usability, human factors, workflow effects, privacy, security, failure recovery, and meaningful human oversight.

Question 9: Would a benchmarking structure such as the one described above be likely to provide adequate evidence of clinical knowledge, analytic capabilities, safety behavior, communication, and generalizability to support a reasonable assurance of safety and

effectiveness? Are there elements that are missing, redundant, or inappropriately categorized? Are there externally developed standards that could be leveraged?

The proposed categories are a strong starting point, but benchmarking should include more than clinical knowledge and output accuracy. FDA should also assess:

- Hallucinations and harmful omissions;
- Calibration and communication of uncertainty;
- Source fidelity and traceability;
- Performance across patient subgroups and care settings;
- Robustness to incomplete, unusual, or adversarial inputs;
- Human-AI team performance;
- Usability, automation bias, and clinician workload;
- Safe failure and escalation behavior; and
- Privacy and cybersecurity.

Benchmark reports should quantify rates of hallucinations and harmful omissions rather than merely record whether they occurred. Results should include denominators, severity, clinical context, uncertainty intervals, subgroup rates, and performance under both routine and adversarial conditions. Rare, high-consequence failures should be reported separately so they are not obscured by aggregate accuracy.

Benchmarks should use independent or sequestered datasets where appropriate, and public summaries should identify the device version, intended use, test population, performance thresholds, and known limitations.

For moderate- and high-risk devices, FDA should encourage qualified independent adversarial testing that probes misleading prompts, unusual inputs, workflow failures, and attempts to bypass safeguards. Common data models, standardized data elements, and consistent reporting formats would make results more comparable across manufacturers, populations, and deployment settings. Benchmarking should be repeatable after material updates and should evaluate the complete deployed system rather than the foundation model alone.

Question 11: CDRH is considering that clinical confirmation for a GenAI-enabled device might not require a prospective clinical study in every case, and has described above a range of approaches of increasing rigor and patient exposure. How might a sponsor select and justify a confirmation approach tailored to a device's intended use and proportionate to the device's risk profile? Are there device types or risk profiles for which one or more of these approaches would be insufficient or inappropriate? How might the anticipated distribution of real-world inputs be taken into consideration? Are there other methods of clinical confirmation that might help inform the evaluation of GenAI-enabled devices?

FDA should use a staged, risk-proportionate approach:

- Lower-risk functions may be supported by retrospective evaluation, usability testing, and representative workflow simulations.
- Moderate-risk functions may require shadow deployment, multi-site testing, clinician adjudication, and subgroup analysis.
- High-risk, autonomous, or action-taking functions may require prospective comparative studies and, when appropriate, randomized trials.

Clinical confirmation should reflect realistic inputs, edge cases, underserved populations, different care settings, and the performance of the human-AI team when the device is intended to assist clinicians. Documentation should be standardized and should avoid duplicating information already maintained through quality-management systems.

Question 13: For which clinical domains, device functions, or subpopulations is synthetic data particularly well-suited, or particularly inadequate, as a supplement to real world evidence? What safeguards would mitigate the risk that synthetic data generated by models of the same class as the device under evaluation reproduces the very performance gaps the evaluation is intended to detect, particularly for underrepresented subgroups?

Synthetic data can supplement real-world data for rare conditions, safety-critical edge cases, and populations that are difficult to sample, but it should not replace real-patient evidence for high-risk uses.

Sponsors should document how synthetic data were generated, demonstrate that clinically important relationships and subgroup characteristics were preserved, and evaluate whether the generating model reproduced existing biases. Final performance claims should be confirmed using independent real-world data, particularly for historically underrepresented populations.

Question 14: For open-ended device outputs, how should performance comparators and acceptance criteria be selected? When a panel of qualified clinicians serves as the comparator, how should the applicable standard (for example, the standard of care versus the performance of a median clinician in practice) be defined and justified? Should generalist or specialist physicians be used as a performance standard? When should human-AI team performance, rather than the device operating alone, serve as the basis for evaluation?

Comparators should reflect the device's intended use and the applicable standard of care. The performance of a "median clinician" should not be the primary benchmark when prevailing practice falls below evidence-based standards.

Generalists should serve as comparators for generalist-facing tools, while specialists should be used for specialty-level tasks. Assistive systems should be evaluated primarily through human-AI team performance, while autonomous systems should also demonstrate acceptable standalone performance. Evaluation should include clinical outcomes, harmful omissions, equity, usability, workload, and resource utilization, not merely agreement with clinicians.

CDRH should not assume that adding a human reviewer improves performance. Studies should compare the unaided clinician, the device alone when relevant, and the human-AI team, and should examine whether human and model errors are complementary or correlated. Human review should be credited as a safeguard only when it measurably improves error detection, clinical decisions, or outcomes under realistic time and workflow constraints.

Question 16: Should independent third parties be involved in some or all aspects of a competency-based approach, including device benchmarking and clinical confirmation? If so, in what ways might qualified, independent third-party participation contribute to this approach, and what qualifications and independence criteria should apply? Are there aspects of the assessment for which third-party involvement would be impractical or inadvisable? What safeguards or program design features would be critical to prevent third-party participation from limiting competition or preventing innovation?

AMIA supports an appropriate role for qualified independent third parties, particularly for adversarial testing, maintenance of sequestered evaluation datasets, review of sponsor-developed benchmarks, and clinical confirmation of higher-risk devices. Independent testing can reduce optimization to known tests and provide greater confidence that performance generalizes beyond the sponsor's development environment.

FDA should establish transparent qualifications, conflict-of-interest protections, standardized methods, data-security requirements, and mechanisms for auditing third-party performance. Participation should be scaled to risk, should not impose disproportionate costs on smaller developers, and should not transfer the manufacturer's ultimate responsibility for safety and effectiveness.

Section VI: Postmarket Monitoring for GenAI-Enabled Devices

Overarching Recommendation to CDRH Regarding Questions 18 to 24

After CDRH clarifies which GenAI-enabled functions fall within the medical-device framework, AMIA recommends that the agency address a related infrastructure need that applies across Questions 18-24. For functions subject to FDA oversight, CDRH should work with other HHS agencies and relevant stakeholders to establish a standardized, interoperable framework for recording AI exposure and provenance.

The framework should allow authorized stakeholders to determine which AI-enabled device or function was used, which product and model version was deployed, which patient and encounter were involved, the clinical context in which the technology was used, and what occurred afterward. These data are necessary to connect real-world use with clinical processes and outcomes as models, data environments, and workflows change.

The framework should be vendor-neutral and should not require clinicians to document each AI interaction or manufacturers to submit every interaction to FDA. Relevant data should be generated automatically wherever possible, remain proportionate to the device's risk, and be available through interoperable formats that existing EHRs, registries, distributed research networks, and other authorized data infrastructures can use.

The Discussion Paper appropriately recognizes that variable outputs and post-deployment changes may limit what premarket evaluation can establish. If FDA relies more heavily on postmarket evidence, standardized exposure and provenance data will be necessary to evaluate safety, effectiveness, generalizability, subgroup performance, and changes over time.

AMIA therefore encourages CDRH to treat this data infrastructure as a core component of the postmarket framework for in-scope GenAI-enabled medical devices and to develop it jointly with ONC, CMS, standards-development organizations, EHR vendors, health systems, manufacturers, patients, clinicians, and researchers. Existing resources, including multi-institutional EHR networks such as Epic Cosmos, illustrate how standardized data generated during routine care could support national-scale evaluation without requiring a new centralized federal database or reliance on a single proprietary platform.

Question 18: CDRH is considering whether it may be appropriate to accept greater premarket uncertainty regarding a GenAI-enabled device's benefit-risk profile through greater reliance on postmarket monitoring. Under what conditions might such an approach be appropriate, and what characteristics of a monitoring program would need to be in place to justify reduced premarket evidence? Are there device types or risk profiles for which this approach would not be appropriate?

AMIA supports greater reliance on postmarket monitoring only in limited circumstances. Reduced premarket evidence may be appropriate for lower- or moderate-risk devices when potential harms are reversible, meaningful human review is available, deployment is controlled, monitoring can detect safety and subgroup concerns, and the manufacturer can rapidly suspend, correct, roll back, or withdraw the product. This approach generally should not apply to autonomous, emergency, diagnostic, treatment-directing, medication-dosing, or other high-consequence functions. Postmarket monitoring should not replace premarket mitigation of serious foreseeable risks.

Any monitoring program used to justify reduced premarket evidence should combine active performance monitoring, independent review, predefined safety thresholds, subgroup analysis, prompt investigation, and enforceable procedures for corrective action. It should also identify which patients were exposed to which device and version, the clinical context in which the device was used, and the subsequent processes and outcomes.

Routine data generally identify medication exposure, and Unique Device Identifiers support identification of many implanted devices. CDRH should consider a comparable structured AI

exposure event for GenAI-enabled medical devices. Depending on the device and its risk, the event should include:

- The patient and encounter;
- The device or software function, manufacturer, and standardized product identifier;
- The deployed model and version, including the underlying foundation-model version when relevant;
- The date, time, clinical setting, and user;
- Whether the function was invoked and whether its output was presented to a clinician or patient;
- The type of output, recommendation, or action;
- The clinician's or patient's response, such as acceptance, modification, or rejection, when relevant; and
- Sufficient provenance to link the exposure to subsequent care processes and outcomes.

The minimum dataset should be proportionate to risk and generated automatically wherever possible so the monitoring system does not create new documentation work for clinicians. This infrastructure would make reliance on postmarket evidence more credible by allowing FDA and manufacturers to determine where, when, and how a device was used and whether outcomes differed after exposure.

Question 19: Please comment on the potential approaches to postmarket performance evaluation, including periodic re-benchmarking, sample-based clinician review, and performance degradation monitoring. What additional approaches should CDRH consider, and how should the cadence and triggering events for reassessment be determined?

AMIA supports periodic re-benchmarking, independent clinician review of real-world samples, performance-degradation monitoring, and adversarial testing as components of a broader algorithmovigilance program. Monitoring should be continuous or sufficiently frequent to detect meaningful change, rather than limited to fixed periodic reviews, because performance may shift with changes in models, data sources, patient populations, connected tools, or clinical workflows.

The program should track clinical outcomes, near misses, harmful outputs, complaints, clinician overrides, care-escalation patterns, subgroup performance, usability and workflow burden, cybersecurity events, and integration failures. Where technically valid, calibrated uncertainty measures may help identify outputs requiring human review, but sponsors should validate those measures rather than present unverified confidence values as safeguards.

CDRH should also incorporate longitudinal real-world outcome monitoring based on standardized AI exposure and version data. These data would support evaluation across institutions and clinical settings, comparisons across patient subgroups, detection of uncommon safety signals, analysis of clinician interaction with AI outputs, and before-and-after assessment of new device or model

versions. Without reliable exposure and version identification, these analyses will remain incomplete.

The monitoring cadence should be risk-based. Reassessment should also be triggered by model updates, new data sources, changes in intended use, deployment in new populations or settings, performance-threshold breaches, adverse events, cybersecurity incidents, and material workflow changes. Findings should feed promptly into investigation, corrective action, controlled validation, and change management. Common data models, interoperable reporting formats, and automated capture should support aggregate learning while minimizing documentation burden.

CDRH should build on existing infrastructure rather than create a disconnected reporting channel. Relevant pathways include eMDR and MAUDE as medical-device report data become available through AEMS, MedWatch and MedSun, and NEST or Sentinel for active surveillance, as appropriate. Because these systems serve different purposes, integration should use common GenAI data elements for product and model version, exposure, clinical context, outcome, and event severity. Spontaneous reports alone cannot estimate incidence; reliable rates require exposure denominators and defined observation windows.

CDRH also should consider an FDA-governed regulatory sandbox or federated registry in which manufacturers, health systems, clinicians, patients, and researchers can test common data elements and monitoring methods and share deidentified or appropriately aggregated safety signals. Governance should protect patient privacy and proprietary information, use role-based access, define how signals are validated and escalated, and publish aggregate findings when appropriate. Participation should supplement, not replace, mandatory reporting or manufacturer accountability.

Question 20: Please comment on whether the proposed approaches to postmarket monitoring can be facilitated by machine-based supervisory agents. What considerations, including the evaluation and reliability of the supervisory agent itself, should CDRH take into account for such an approach?

Machine-based supervisory agents may help CDRH and manufacturers monitor GenAI-enabled devices by reviewing high volumes of interactions and detecting changes in utilization, output patterns, clinician acceptance or override rates, subgroup performance, or outcomes after a model update. Their usefulness will depend on standardized, reliable data that identify AI exposures, versions, clinical context, and subsequent outcomes.

A supervisory agent should be evaluated as a safety-relevant system in its own right. Validation should address its intended monitoring task, input-data quality, sensitivity to meaningful safety signals, false-positive and false-negative rates, subgroup performance, drift, cybersecurity, and performance under unusual or adversarial conditions. Monitoring plans should include audit logs, transparent alert thresholds, human review and escalation procedures, periodic revalidation, and change control for the supervisory agent.

Supervisory agents should facilitate, not replace, human monitoring programs, validated source data, manufacturer investigation, clinical judgment, or FDA oversight. A supervisory agent should flag and prioritize potential signals for accountable human review rather than make final safety or regulatory determinations. It also cannot compensate for the inability to identify which device and version was used or to link that exposure to clinical outcomes.

Question 21: What roles might clinicians, healthcare institutions, professional societies, standards-setting bodies, and other stakeholders appropriately play in postmarket monitoring, and how can these roles be structured without diffusing manufacturer accountability?

Postmarket monitoring should use shared infrastructure and clearly assigned roles, but the manufacturer must remain accountable for the device's safety and effectiveness. FDA should not permit contracts to transfer the manufacturer's core monitoring, investigation, reporting, or corrective-action responsibilities to clinicians or health systems.

An appropriate division of responsibility would include:

- Manufacturers should provide persistent identifiers for products, models, and versions and remain responsible for monitoring, investigation, regulatory reporting, and corrective action;
- Health care institutions and EHR vendors should oversee local implementation, integration, and training; detect site-specific problems; and automatically capture standardized AI exposure data;
- Clinicians and patients should be able to report safety and usability concerns through streamlined, accessible mechanisms that do not create excessive administrative burden;
- Professional societies should help develop clinical competencies, benchmarks, reporting standards, and education;
- Standards-development organizations, research networks, and registries should establish and use common representations for AI products, versions, exposures, outputs, provenance, and safety events; and
- HHS agencies should coordinate their authorities: ONC should examine certification criteria, interoperability standards, and APIs; CMS should consider appropriate quality, health IT, or payment-program mechanisms; and FDA should define the information manufacturers must provide and how real-world data may support regulatory evidence.

Standardizing the underlying data, rather than mandating a single repository, would allow existing EHR networks, registries, distributed research networks, and other real-world data infrastructures to support multi-institutional evaluation. Epic Cosmos is one illustrative example, but the policy should remain vendor-neutral. Linking AI exposure with the device, model, version, clinical context, user interaction, subsequent care, and patient outcome could support subgroup analyses and detection of uncommon safety signals without requiring a new centralized federal database.

This approach recognizes that the data infrastructure is an ecosystem responsibility while preserving clear manufacturer accountability for each regulated device. Any shared registry or sandbox should use role-based access, privacy and cybersecurity controls, protections for trade secrets and confidential commercial information, transparent criteria for escalating safety signals, and public reporting of appropriately aggregated findings. Participation should not replace mandatory reporting obligations.

Question 22: CDRH envisions that a manufacturer’s premarket competency-based assessment might serve as a baseline against which post-deployment modifications could be re-evaluated. How might the extent of re-benchmarking or other evidence be scaled to the nature and expected impact of a given modification? For example, are there categories of change that might not significantly affect safety or effectiveness of a GenAI-enabled device, or might be appropriately managed within a sponsor’s quality management system versus requiring FDA premarket review and authorization before implementation? Are there categories of changes that might be appropriate for inclusion in a PCCP?

CDRH should use a graduated change-control approach based on the likelihood that a modification could alter clinical function, performance, user interaction, or risk. Nonmaterial changes that do not affect clinical logic, model behavior, or validated interfaces may be managed within the manufacturer's quality management system when supported by verification and a documented rationale. Intermediate changes should receive targeted revalidation and any required notification, while material changes should receive premarket review unless they fall within an authorized Predetermined Change Control Plan.

Changes to model weights or architecture, training data, input sources, output generation, safeguards, intended use, target population, autonomy, or connected tools should trigger targeted re-benchmarking and, where appropriate, clinical confirmation. A manufacturer's characterization of a change as an improvement should not determine the review pathway; the pathway should depend on demonstrated effects on safety, effectiveness, intended use, and foreseeable failure modes.

The premarket competency assessment should serve as the baseline for each comparison. Every deployed version should have a persistent identifier, and standardized exposure data should support before-and-after analysis of real-world performance and outcomes. A PCCP may be appropriate for reasonably foreseeable categories of change when validation methods, acceptance thresholds, monitoring requirements, and rollback procedures are prespecified.

Question 23: How might PCCP concepts or other change-control approaches be adapted for GenAI-enabled devices when the nature or scope of future modifications cannot be fully prespecified?

When exact modifications cannot be fully prespecified, a PCCP should define bounded categories of permitted change and prespecify the process for assessing their effects. The plan should identify

excluded changes and require additional FDA review for modifications that alter intended use, materially expand autonomy or risk, exceed defined performance limits, or otherwise fall outside the authorized scope.

At a minimum, the PCCP should specify data sources, validation methods, clinical and subgroup acceptance criteria, human-factors and cybersecurity checks, documentation requirements, deployment controls, post-deployment monitoring, notification and reporting, and stop or rollback criteria. Within an authorized PCCP, FDA should permit an expedited pathway for an update that maintains or improves safety and effectiveness when the sponsor meets the prespecified criteria, demonstrates that the update does not introduce a new material failure mode, assigns a persistent version identifier, and maintains monitoring and rollback capability. Updates outside those boundaries should follow the ordinarily applicable premarket review pathway.

Question 24: For devices built on third-party foundation models, changes to the underlying model may be initiated by the third-party model developer rather than the device manufacturer. How can a manufacturer detect, evaluate, and respond to such changes in a timely manner? What mechanisms—for example, contractual, technical, or through a PCCP—could provide reasonable assurance that third-party developer-initiated changes do not compromise the safety or effectiveness of the device?

The device manufacturer should remain responsible for detecting, evaluating, and responding to changes made by a third-party foundation-model developer. Contracts and technical controls should provide:

- Advance notice of planned and emergency model changes;
- Persistent version identification and, where feasible, version pinning;
- Detailed release notes and agreed criteria for material changes;
- Access to evaluation results, known limitations, incident information, and relevant audit records;
- Automated regression testing and risk-based validation before a changed model is deployed in the device;
- Rapid suspension, rollback, and safe fallback capabilities; and
- Immediate notification of safety-related failures or vulnerabilities.

Exposure records should identify both the commercial device and the relevant underlying model or configuration when technically feasible and relevant to safety or effectiveness. This provenance would allow manufacturers and regulators to determine whether a change in real-world performance coincided with a foundation-model update.

A PCCP may address reasonably foreseeable third-party updates, but it should not permit automatic deployment solely because the third-party developer characterizes an update as an improvement. The manufacturer should validate the complete device after any material change, including its safeguards, workflow integration, and connected tools. An update within authorized

boundaries may use the expedited pathway described above; a material or out-of-scope change should receive the ordinarily applicable review. Foundation Model Device Master Files could support timely information exchange, but they should not replace contractual notice, technical controls, or the manufacturer's responsibility for continued safety and effectiveness.

Section VII: Other Topics

Question 25: Would voluntary Foundation Model MAFs be practical for foundation model developers to provide and to keep current, and what content might they include? Given that participation would be voluntary and that model developers may have limited incentive to disclose safety-relevant information, what would make such a program sufficiently useful for premarket review? Are there alternative mechanisms CDRH should consider for obtaining information about underlying foundation models?

AMIA supports Foundation Model Master Files because they could reduce duplicative submissions and improve FDA's understanding of models used across multiple devices. However, a purely voluntary program may provide incomplete information, particularly for safety-critical models.

Files should use standardized, machine-readable templates and include model architecture, supported uses, training-data provenance, known limitations, failure modes, subgroup evaluations, guardrails, version history, cybersecurity information, update-notification commitments, and audit-log capabilities. FDA should protect proprietary details while requiring an accessible public summary of safety-relevant information.

Question 26: Are there additional considerations that inform the premarket and postmarket evaluation of agentic GenAI-enabled devices, beyond those applicable to non agentic GenAI-enabled devices? How should the elevated risk associated with autonomous multi-step action, tool use, and reduced opportunity for human review be reflected in acceptance criteria and oversight?

Agentic devices require heightened oversight because errors can compound across multiple steps before a clinician can intervene. Evaluation should address:

- The device's permissions and access to external systems;
- The safety of tool and API use;
- End-to-end task completion, rather than isolated outputs;
- Error propagation across sequential actions;
- Human authorization checkpoints;
- Stop, override, and rollback mechanisms;
- Failure containment and safe handoff;
- Immutable audit trails; and
- Performance under adversarial, incomplete, and unexpected conditions.

A nominal “human in the loop” should not be treated as an adequate safeguard unless the person has sufficient time, information, authority, and expertise to meaningfully review and stop the device’s actions. FDA should increase evidentiary and monitoring requirements as autonomy, action complexity, and potential harm increase.

AMIA appreciates the opportunity to provide these comments and welcomes continued engagement with CDRH. We encourage FDA to pursue a clear, risk-proportionate, and lifecycle-based approach that protects patients while supporting responsible innovation. Please contact Tayler Williams, Senior Manager of Public Policy, at twilliams@amia.org with any questions.

Sincerely,



Eileen Koski
Chair, AMIA Public Policy Committee