

October 16, 2026

Rick Abramson, MD
Director
Digital Health Center of Excellence
Center for Devices and Radiological Health
U.S. Food and Drug Administration
10903 New Hampshire Avenue
Silver Spring, MD 20993

Re: Considerations for the Regulation of Generative AI-Enabled Medical Devices: Discussion Paper and Request for Feedback

Dear Director Abramson:

On behalf of the physician and medical student members of the American Medical Association (AMA), I appreciate the opportunity to provide feedback to the U.S. Food and Drug Administration (FDA) on its Discussion Paper and Request for Feedback regarding [Considerations for the Regulation of Generative Artificial Intelligence \(GenAI\) Enabled Medical Devices](#). Given the rapid pace of technological innovation in healthcare, it is imperative that we pursue new and innovative regulatory approaches that better meet the needs of algorithmic-enabled medical devices. The AMA commends FDA for initiating this early dialogue and appreciates FDA's thoughtful approach as laid out in this discussion paper. We look forward to continuing to engage with FDA and other interested parties as we work to ensure the safety and performance of GenAI-enabled technologies.

The integration of AI into clinical care has the potential to improve diagnostic accuracy, reduce time to diagnosis, enable more personalized treatment recommendations, and support patients in managing and understanding their health. Realizing these benefits, however, depends on establishing an appropriate regulatory framework that ensures AI-enabled technologies are safe, effective, and reliable for use in clinical care.

The AMA recognizes that AI technologies present novel challenges for FDA oversight that may not be fully addressed by regulatory approaches developed for static medical devices. We therefore support FDA's efforts to evaluate and develop regulatory approaches that are appropriately tailored to the unique and evolving characteristics of AI-enabled technologies while maintaining robust standards for patient safety and product performance. Consistent with the AMA's longstanding position, such oversight should incorporate a least-burdensome, risk-based framework that supports continued innovation in healthcare.

The AMA offers the following perspectives:

Overarching principle: Regulatory agility cannot come at the expense of FDA's reasonable assurance of safety and effectiveness.

We strongly support FDA's commitment to a nimble, least-burdensome, total product lifecycle approach to generative AI-enabled (GenAI) medical devices. These principles are particularly important for technologies that will evolve rapidly and for which traditional evaluation methods may not always be fit for purpose.

At the same time, the speed and complexity of technological change should not lower the evidentiary bar for establishing reasonable assurance of safety and effectiveness before a product reaches patients and clinicians. Difficulty evaluating a technology is not itself a justification for accepting inadequate evidence. Nor should regulatory agility result in patients, clinicians, or health systems effectively becoming the final layer of safety and effectiveness evaluation after deployment or result in them bearing risks that should have been more fully characterized before market authorization.

We recognize that the characteristics of GenAI-enabled devices may make exhaustive premarket evaluation challenging. As FDA notes, devices with broad intended uses or open-ended inputs and outputs may make it unreasonable to evaluate every conceivable input a device could encounter after deployment. The inability to evaluate every conceivable input, output, or interaction, however, should not be conflated with an inability to establish sufficient evidence of safety and effectiveness for the device's intended use before authorization and deployment. Premarket evaluation should characterize the bounds of performance, uncertainty, and foreseeable failure modes sufficiently to support safe use in the intended population and setting. Where residual uncertainty remains, post market monitoring can provide an important complement to premarket evidence, but it should not substitute for appropriately rigorous premarket evaluation nor make patients and clinicians responsible for detecting inadequately characterized risks or determining whether a device is safe and effective through their use of it.

A total product lifecycle (TPLC) approach should allow the type and distribution of evidence to evolve, including greater use of real-world evidence and continuous performance monitoring, without transferring the fundamental burden of demonstrating safety and effectiveness away from the manufacturer. post market evidence should complement, rather than substitute for, an appropriately rigorous premarket demonstration of safety and effectiveness. Premarket evaluation should establish sufficient confidence for the proposed intended use and deployment context before exposure; post market monitoring should then confirm, characterize, and continuously reassess performance in real-world use. FDA's own framing that premarket evidence may be "complemented by post market evidence" provides a strong foundation for this approach.

Overarching principle: Physician expertise should inform the standards, not serve as the safety net. Physicians have an essential role across the TPLC to ensure that GenAI-enabled medical devices are safe, effective, and clinically useful. That role should be concentrated where physician expertise adds the greatest value: defining clinically appropriate performance standards, informing representative evaluation, adjudicating clinically consequential outputs, and shaping fit-for-purpose approaches to ongoing clinical surveillance.

Professional medical organizations can provide an important bridge between regulators, manufacturers, and clinical practice by helping establish rigorous and reproducible methods for incorporating physician expertise throughout the total product lifecycle. The AMA welcomes the opportunity to work with FDA and other interested parties to develop these approaches, including methods for clinician adjudication, sample-based clinical review, identification of clinically consequential performance signals, and communication of emerging concerns to physicians.

At the same time, engaging physicians in the evaluation and surveillance ecosystem should not transfer responsibility for product performance to individual clinicians. Physician expertise should help establish

and evaluate the standards GenAI-enabled medical devices are expected to meet; physicians should not become the uncompensated safety infrastructure required to make inadequately performing technologies safe in practice. Manufacturers remain responsible for demonstrating and maintaining the safety and effectiveness of their products, including resourcing any clinician participation necessary to meet those obligations.

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.

The two-axis risk-based framework is a useful starting point, and increasing independence of device activity is generally associated with increasing risk. However, engagement with the device by a licensed healthcare professional (HCP) should not be presumed to mitigate device risk. Where HCP supervision is relied upon as a risk control, its effectiveness and feasibility should be demonstrated as part of the evaluation of the device in its intended use environment.

FDA's framework currently treats a function acting with continuous HCP supervision as meaningfully different from one acting fully autonomously. This distinction is reasonable only where HCP supervision provides a meaningful and effective safeguard. The presence of an HCP in the workflow, or a requirement that an HCP review or approve an AI-generated action, does not by itself establish that the HCP can reliably identify and prevent an unsafe or ineffective action.

We support FDA's proposed consideration of both the consequences of an incorrect output and the degree to which a GenAI-enabled device independently directs or takes action. However, we encourage FDA to further examine how it treats HCP supervision within the activity axis, specifically an HCP that is designated as the intended user by the manufacturer.

This assessment should include whether the intended user:

- has access to sufficient transparency regarding the output, its basis, relevant source information, uncertainty, and known limitations to conduct meaningful review;
- has the clinical competency necessary to independently evaluate the output and recognize an error, including where specialist knowledge may be required;
- has received any education or training necessary to supervise the device safely;
- has sufficient time and cognitive capacity within the intended clinical workflow to conduct the expected review;
- has meaningful authority and opportunity to interrogate, modify, reject, or stop the recommended action, rather than serving as a nominal approval step in a substantially automated process; and
- has appropriate resources and reimbursement for the work and time required to provide the expected level of oversight.

The AMA believes that explainable AI with safety and efficacy data should be the expected form of AI tools for clinical applications with very rare exceptions. Explainability requires an AI device's explanation of how it arrived at its output to be interpretable and actionable by a qualified human; it should not be used as a substitute for other means of establishing safety and efficacy of AI tools, such as through randomized clinical trials.

FDA should also consider the potential for automation bias and over-reliance that the technology may promote when assessing HCP supervision as a safeguard. As GenAI-enabled devices become increasingly capable and routinely produce appropriate outputs, clinicians may become less likely to identify infrequent but clinically consequential errors. Effective supervision therefore cannot be inferred from the HCP's underlying clinical competency alone. The risk that a particular technology may be prone to promoting automation bias or over-reliance in any intended user should be integrated into the risk assessment process.

Further, HCP supervision may introduce risks into the clinical system. Reviewing and independently validating AI-generated outputs requires clinician time and cognitive resources and may contribute to cognitive burden, alert fatigue, workflow fragmentation, or displacement of other clinical activities. Assessment of HCP supervision as a risk mitigation measure should consider its impact on the safety of the human-AI system as a whole, including the HCP's ability to detect errors, the resources required for effective oversight, and any new risks introduced by the supervisory task itself.

This is particularly important where HCP supervision is relied upon to support a lower-risk classification or reduced evidentiary burden. FDA should avoid a regulatory construct in which the manufacturer's regulatory burden is reduced because an HCP has been inserted into the workflow, while clinicians and healthcare systems absorb the additional investment, workload, and risk required to make that supervision effective. If HCP oversight is necessary to achieve an acceptable benefit-risk profile, it should be treated as a required condition of safe use. The manufacturer should be required to demonstrate that the proposed oversight can be performed safely and reliably in the intended use environment and provide the device-specific information, training, and other supports that are necessary to enable it.

This framing is also consistent with FDA's recognition elsewhere in this section that some device outputs may not be independently assessable by the user, and that the ability of an HCP to independently evaluate an output may depend on whether the HCP possesses relevant specialist expertise. These considerations should be incorporated into FDA's treatment of HCP supervision on the activity axis rather than addressed only as separate considerations.

In addition to considerations regarding the level at which licensed healthcare professionals are expected to engage with GenAI-enabled devices, FDA should consider what additional elements could ultimately impact the clinical risk of the tool. Modifiers such as reversibility, downstream safeguards, time sensitivity, explainability, and transparency, as well as any context- or system-specific factors, should impact a device's placement within the risk matrix and should impact evidentiary and monitoring requirements. Licensed clinician oversight should not be considered an appropriate substitute for regulatory oversight when, for example, actions are irreversible or the device is used in time-sensitive situations.

Additionally, entities involved in the risk assessment will be critically important to ensuring safety and performance of technologies under FDA review. Risk assessments provided by the manufacturers, alone, should not be considered reliable and should undergo additional review and approval by licensed professionals with experience in the domain in which a GenAI-enabled device is intended to be used. Manufacturers may be prone to downplaying the risk level of their products in order to lower evidentiary burdens, which will increase downstream risk levels for physicians and their patients. FDA should consider what appropriate mechanisms may exist for appropriate independent and reliable risk-assessment.

Center for Devices and Radiological Health (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 informational functions while recognizing that directiveness may exist along a continuum rather than as a binary distinction?

We agree with FDA that the degree to which an informational function is action-directing exists on a spectrum and should be assessed based on the substance and context of the output rather than particular words or phrases submitted by manufacturers for marketing. We particularly support FDA’s recognition that inclusion of statements such as “talk to your doctor” or “I am not a medical professional” do not make otherwise action-directing information less directive. Likewise, how manufacturers choose to describe the intended use of their products has become a way to evade higher evidentiary burdens despite how the product actually functions and the true risk associated with the product’s outputs. Risk assessments, and therefore regulatory status, should not rest on submitted marketing language alone and should be determined by the actual function of the device.

Available evidence suggests that generic disclaimers should not be presumed to meaningfully mitigate the risk associated with an otherwise action-directing output. Recent experimental research examining AI-generated medical advice found mixed effects of disclaimers on users’ willingness to seek a second opinion and concluded that disclaimers alone may not reliably prevent over-reliance.¹ Other experimental research on AI-generated information has similarly found inconsistent effects of disclaimers on users’ perceptions of content and source credibility.²

As such, FDA should assess directiveness based on the overall interaction and the action a reasonable user is likely to take in response to it. Disclaimer language should not be assumed to move an otherwise action-directing function into a lower-risk category. Where a manufacturer proposes a disclaimer or other risk communication as a substantive risk mitigation measure, the effectiveness of that intervention should, as in other cases, be supported by evidence in the intended user population and use context.

Additionally, FDA should mandate that manufacturers of GenAI-enabled medical devices clearly indicate the intended uses, intended users, intended populations, limitations, performance characteristics, and validated uncertainty measures. FDA should consider how manufacturers should document where a given output type is meant to be placed along this continuum and could provide examples demonstrating how different factors may cause a device or device function to move along the spectrum. Manufacturers should test specifically for unintended drift towards directiveness.

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 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?

¹ Denecke K, Gabarron E. “‘But it sounded confident’: the role of accuracy, tone, and disclaimers in users’ medical decision-making,” *Front. Artif. Intell.* 2026;9:1785085. doi: 10.3389/frai.2026.1785085

² Lermann Henestrosa A, Kimmerle J. “‘Always check important information!’ - the role of disclaimers in the perception of AI-generated content.” *Comput Hum Behav Artif Hum.* 2025;4:100142. <https://doi.org/10.1016/j.chbah.2025.100142>.

We encourage FDA to frame this question primarily around demonstrated performance in the intended user population and use environment. Patient-facing use may introduce important considerations for evaluation, but patient status itself should not serve as a proxy for increased risk. A risk assessment of patient-facing technologies should focus on the design of the device and the consequence of the function rather than assumptions that patients are unable to understand or unable to act on the provided information.

Manufacturers should demonstrate that a device performs safely and effectively when used by its intended users under realistic conditions, including whether users understand its outputs, respond appropriately to uncertainty, recognize when additional care is needed, and take appropriate downstream actions. Patient-facing technologies should clearly disclose that the outputs are AI-generated and should be required to clearly and meaningfully disclose that outputs are generated solely on what the individual shared and may not be complete. They should explicitly and persistently remind users that they are not a substitute for a physician or other licensed healthcare professional.

This approach would preserve FDA's risk-based framework while avoiding assumptions about patient capability that could unnecessarily restrict access to technologies capable of safely expanding access to clinical information and care.

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?

We encourage FDA to distinguish between clinical expertise that defines the intended use environment and clinical expertise that is relied upon to compensate for inadequate device performance. If safe use depends on a specialist recognizing and correcting errors that a generalist would not reasonably detect, the device may not have demonstrated sufficient safety and effectiveness for use by generalist HCPs.

The appropriate question is therefore whether the GenAI-enabled device demonstrates adequate performance when deployed with its intended users and under clinically representative conditions. A device intended to extend specialist knowledge to generalist HCPs should be evaluated specifically in that context, including evaluation of its performance with generalists and the performance of the resulting human-AI system. Foreseeable circumstances in which a generalist might misinterpret or misapply an output designed for more specialized expertise should also be included as part of the evaluation process. FDA itself recognizes the potential for these tools to extend access to specialty knowledge. The evidentiary framework should determine whether they can do so safely and effectively.

Where specialist expertise is a necessary condition for safe use, that limitation should be reflected in the intended use, intended user, and evaluation of the device. Expertise required merely to detect and correct otherwise unacceptable device errors should not be treated as a substitute for adequate product performance.

FDA should consider what additional safeguards may be put into place when considering technologies that may expand generalists' reach into specialty contexts. These safeguards could include specialty-specific validation, criteria for specialist consultation, and escalation mechanisms for cases outside the device's validated scope.

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?

For multi-turn conversational GenAI-enabled devices, FDA should assess risk across realistic conversational trajectories and characterize the intended use based on the highest-risk function the device can reasonably be expected to perform within its intended use.

If a device can progress from providing non-directive information to providing action-directing information during reasonably foreseeable interactions, its evaluation should account for that action-directing capability. For example, a conversational device may initially provide general information about the side effects of a medication, but over the course of an exchange, progress to recommending that a patient stop taking the medication or change the dose based on reported symptoms. The latter function is action-directing and should inform the device’s risk assessment and evaluation, even if most interactions remain informational. The frequency with which the higher-risk behavior occurs may be relevant to overall benefit-risk assessment, but should not result in the device being characterized according to its more common, lower-risk behavior.

Manufacturers should therefore demonstrate performance across representative conversational trajectories, including those that reach the highest-risk functions within the device’s intended or reasonably foreseeable use.

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?

The AMA generally supports FDA’s exploration of a competency-based approach combining device benchmarking with clinical confirmation. For GenAI-enabled devices with broad inputs, open-ended outputs, and performance that cannot reasonably be characterized through exhaustive testing, a competency-based approach may offer a practical and scientifically rigorous path to establishing reasonable assurance of safety and effectiveness.

However, the clinician analogy underpinning this framework carries an essential implication that should be incorporated explicitly into the regulatory model: competency and accountability must travel together. As FDA’s discussion acknowledges, clinicians operate within systems of professional, legal, and reputational accountability when their performance falls below expected standards. A regulatory framework modeled in part on clinician competency cannot adopt the flexibility of that model while leaving its corresponding accountability mechanisms unresolved.

Manufacturers should remain accountable for the safety and effectiveness of their devices across the total product lifecycle, including performance that emerges through interaction with users and real-world environments. Neither the probabilistic nature of GenAI outputs nor the impossibility of testing every conceivable scenario should diffuse that accountability across clinicians, patients, deploying institutions, foundation-model developers, or other participants in the ecosystem.

This becomes especially important when a device operates alongside a clinician. Clinician involvement should not become the mechanism through which accountability for device performance migrates from the manufacturer to the HCP. Where safe and effective performance depends on particular conditions of use, including HCP oversight, training, workflow design, or other safeguards, the manufacturer should demonstrate that those conditions are feasible and effective in the intended deployment environment.

Clinical confirmation is therefore critical to establishing that performance demonstrated through benchmarking translates into safe and effective performance in the intended clinical environment. FDA appropriately recognizes that benchmarking alone may not capture interactions among the device, users, workflows, and patient populations in actual clinical use. It is critical that evaluation establishes the performance of the device as deployed, including the human-AI system where applicable, while maintaining clear accountability for the regulated product.

How might the two-axis risk framework described in Section IV be considered within the competency-based approach described above to help determine the level of evidence needed for a premarket submission?

The two-axis risk framework could provide a useful basis for scaling the nature, rigor, and amount of evidence required within a competency-based approach. As the independence of a device function or activity increases, or the potential negative consequences of an incorrect output increase or intensify, the evidentiary burden for manufacturers should increase accordingly. This principle should apply regardless of whether a device's activity or function has been pre-characterized as informational in nature.

FDA should exercise particular caution in using the presence of HCP supervision to support a lower evidentiary burden. As discussed in response to Question 1, HCP involvement should only be treated as a risk mitigation measure where there is evidence that the proposed supervision can be performed effectively by the intended user in the intended use environment. The presence of an HCP in the workflow should not compensate for inadequate device performance or become a basis for authorizing technology that requires clinicians to routinely identify and correct otherwise unacceptable errors. Where HCP supervision is necessary to achieve an acceptable benefit-risk profile, the competency-based assessment should therefore evaluate the human-AI system under clinically representative conditions, including the HCP's ability to recognize errors, the information and training available to support review, the time and workflow required for effective oversight, and the performance of the combined system. The evidentiary burden should reflect the performance the device must achieve for its intended use, rather than an assumption that clinicians will backstop performance deficiencies after deployment.

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?

Performance comparators should be selected based on the device's intended use and should reflect the realities of contemporary clinical practice. Where FDA evaluates the performance of a human-AI team, the conditions under which the HCP interacts with the technology should be representative of the conditions under which the device will actually be used. A fundamental guiding principle should be that safety is a *system* property, rather than a *software* (or in this case, *GenAI-enabled software*) property. This implies that GenAI-enabled functions do not exist in a vacuum and their actual performance in any given system, including their safety and effectiveness, cannot be fully gauged until they have been implemented and validated within the actual system and settings and among the end-users and patients for which they will be deployed.³

³ Leveson N. *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press; 2012.
doi.org/10.7551/mitpress/8179.001.0001.

Physicians currently operate under significant time and workflow constraints and may not have dedicated time, specific training, or adequate reimbursement to independently interrogate every questionable AI-generated finding. An evaluation that provides clinicians with additional time, information, training, or opportunities to investigate outputs beyond those reasonably available in practice could substantially overestimate the performance and safety of the human-AI system after deployment.

Clinical studies and other performance evaluations should therefore reproduce expected conditions of use, including the time available for review, relevant training, information presented by the device, workflow integration, and expectations for physician oversight. If safe performance requires resources or competencies beyond those ordinarily available in the intended deployment setting, those requirements should be explicitly characterized as conditions of safe use and addressed in the device's evaluation and implementation.

The selection of generalist versus specialist physicians as comparators should similarly follow the intended use and deployment context. A device intended for use by generalist physicians should demonstrate adequate performance in that setting and with those users. Where a device is intended to extend access to capabilities or knowledge that would otherwise require specialist care, evaluation should determine whether it can safely and effectively do so when used by the intended generalist HCP. Specialist performance may provide an informative benchmark, but specialist expertise should not be assumed to be available after deployment to identify and correct device errors.

More broadly, FDA should avoid comparator designs that establish performance under conditions more favorable than those the device will encounter in practice. The relevant standard should reflect how the product and its intended users will perform together in the environment in which care is actually delivered.

Are there ways in which performance might be assessed relative to the care, technology, or course of action likely to occur in the absence of the device, rather than to the comparators described in this section? What approaches might be used to identify and justify a comparator such as unaided clinical judgment, delayed specialist review, or no intervention?

We recognize the important opportunity for GenAI-enabled medical devices to expand access to clinical capabilities where appropriate care is unavailable, delayed, or difficult to obtain. The availability of alternative care is, therefore, a legitimate consideration in assessing the benefits and risks of a device. FDA should exercise significant caution, however, before allowing limitations in access to care to justify lower standards of device performance. Such an approach could unintentionally institutionalize a two-tier standard of care in which populations already experiencing barriers to healthcare are offered technologies that would not otherwise satisfy expectations for safety and effectiveness. This could easily exacerbate existing healthcare disparities and run afoul of the principles of the Medical Code of Ethics, which states that medicine has an ethical responsibility to promote equitable care for all.⁴

The relevant comparison may appropriately include the care that patients can realistically access, rather than an idealized alternative that is unavailable to them. That comparison should inform assessment of the device's potential benefit; it should not automatically lower the standard for establishing that the device itself is safe and effective for its intended use.

⁴ AMA. Code of Medical Ethics: Opinion 11.2.7 - Responsibilities to Promote Equitable Care. AMA Code of Medical Ethics. <https://code-medical-ethics.ama-assn.org/ethics-opinions/responsibilities-promote-equitable-care>.

This distinction is particularly important as limited access is often concentrated among rural, low-income, or otherwise underserved populations. A regulatory framework should create incentives to develop technologies capable of extending high-quality care to these communities, rather than creating a pathway through which lower-performing technologies become acceptable because existing healthcare delivery has failed to provide adequate access.

Where FDA determines that unmet need or lack of alternatives warrants greater tolerance for uncertainty or different performance expectations, strong guardrails should apply, including a clearly defined population and context of use, transparent characterization of the performance tradeoff, evidence that the device produces a favorable benefit-risk profile in that population, appropriate post market monitoring, and reconsideration of the standard as better alternatives become available.

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?

Clinician adjudication can play an important role in the evaluation of GenAI-enabled medical devices, particularly where outputs are open-ended, context-dependent, or subject to legitimate differences in clinical judgment. Practicing physicians can contribute expertise in defining clinically appropriate performance standards, developing representative clinical scenarios, assessing the clinical significance of errors and variation in outputs, and determining whether device performance would support safe and effective use in practice.

Clinician adjudicators should be selected to reflect the expertise relevant to the device's intended use, including the clinical setting and intended users. Evaluation methods should account for legitimate variation in clinical practice and should distinguish disagreement reflecting reasonable clinical judgment from errors that could affect patient care.

Importantly, clinician expertise is most valuable when it is used upstream to establish what acceptable performance looks like and to evaluate whether a product meets that standard before and across deployment. It should not become a substitute for adequate evidence of device performance or create an expectation that individual physicians will identify and correct inadequately characterized device errors during routine clinical use.

Regardless of clinical adjudication, disinterested, independent third parties should be involved in the validation of benchmark designs, assessment of performance in subgroups, adjudication of open-ended outputs, and confirmation of sponsor- or manufacturer-reported claims or results. To avoid thwarting competition, the FDA should establish a rubric with common criteria and methodologies that are transparent and do not have inequitable effects for smaller or less-resourced manufacturers, along with transparent review and appeal processes.

Professional medical organizations and their members can play an important role in developing best practices for clinician adjudication, including adjudicator qualifications and selection, representative clinical scenarios, methods for resolving disagreement, and criteria for determining the clinical significance of errors. The AMA would welcome the opportunity to work with FDA and other stakeholders to develop these approaches and ensure they are rigorous, reproducible, and grounded in the realities of clinical practice.

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 post market 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?

We do not support reducing the level of premarket evidence necessary to establish reasonable assurance of safety and effectiveness based on the assumption that uncertainty can be addressed through post market monitoring.

Robust post market monitoring will be particularly important for GenAI-enabled medical devices because performance may change or degrade over time and risks may emerge under real-world conditions that could not reasonably have been fully characterized before deployment. FDA appropriately recognizes an important role for post market evidence across the total product lifecycle.

However, accepting greater uncertainty about safety and effectiveness at the time of authorization means that patients and HCPs become the population exposed to that unresolved risk while the evidence necessary to characterize it is generated. This is particularly concerning where failures may cause clinically consequential and potentially irreversible harm or where users cannot readily recognize an incorrect output. Post market monitoring identifies harms after they occur. Our ultimate goal should be to prevent harm from occurring in the first place.

Premarket evaluation need not eliminate all uncertainty or test every conceivable interaction. It should establish sufficient evidence to provide reasonable assurance of safety and effectiveness for the intended use, intended users, and intended system and setting, before deployment. Post market monitoring should then identify emerging risks, characterize performance under real-world conditions, detect degradation, and confirm that the benefit-risk profile remains acceptable over time. Manufacturers should be required to take timely, predefined action when monitoring identifies performance degradation, new or increased risks, or other evidence that the device may no longer meet the conditions under which its safety and effectiveness were established, including corrective action, modification, notification of users, or withdrawal from use as appropriate.

Additionally, FDA is aware of broad concerns regarding liability exposure for physicians when using AI-enabled devices and other AI-enabled systems. When physicians perceive the risk of liability exposure is high, it serves as a potentially significant disincentive to adopt new technology. The suggestion that rigorous post market monitoring can substitute for appropriate premarket evaluation would lend itself to the idea that physicians would be engaging in tools where the risks of patient harm were not fully mitigated prior to deployment. This would arguably amount to a breach in the duty and standard of care owed to patients and could amount to medical negligence, potentially exposing physicians to civil and even criminal liability. The burden of risk identification and mitigation would potentially be shifted to the physician, which represents an unacceptable laundering of liability away from the developer and onto the party that is least able to identify performance issues and mitigate risks to patients. When patients are harmed, the damage does not lessen because FDA lowered its evaluation criteria—it simply attempts to shift the liability exposure away from the manufacturer and towards the remaining parties, meaning the physicians and other clinician end-users.

Please comment on the potential approaches to post market 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?

Post market monitoring for GenAI-enabled medical devices should include ongoing, sample-based clinician review of outputs generated in real-world use, particularly for devices producing clinically consequential or open-ended outputs.

Automated monitoring, periodic benchmarking, and quantitative measures of performance are important components of a post market surveillance system, but may not capture clinically significant changes in the quality, appropriateness, or context of GenAI-generated outputs. Factors that are particularly important to monitor and track for AI/ML-enabled devices include concept drift, covariate shift, and algorithmic stability.⁵ Sample-based review by appropriately qualified clinicians in real-world contexts provides an essential mechanism for identifying emerging failure modes, evaluating whether outputs remain clinically appropriate, and detecting performance degradation that may not be apparent through aggregate metrics alone. Such reviews should oversample safety-critical interaction types, such as those involving escalations or overrides, rather than relying on a random sample to ensure that rare, high-consequence failures are included in reviews.

The responsibility and costs for establishing and resourcing this clinical review should rest with the manufacturer. Manufacturers should be responsible for assembling appropriately qualified clinicians, compensating them for their expertise and time, establishing fit-for-purpose sampling and review methodologies, and acting on findings. Physicians using the device in routine clinical practice should not be expected to absorb an unfunded post market surveillance function on behalf of the manufacturer.

Reassessment should take place at fixed intervals based on the device's risk, level of clinical consequence, degree of autonomy, frequency of use and observed change, and the nature of the model and deployment environment, as well as when certain triggering events occur, such as suspected performance drift, changes in patient population or practice settings, model or workflow updates and changes to underlying foundational models (including those that are seemingly minor or initiated by the sponsor), new clinical guidelines and standards of care, high override rates, as well as any safety events, near misses, or user-reported concerns. The reassessments should be designed to detect both systematic performance changes and clinically important low-frequency failures or deficiencies. Findings should feed into predefined processes for investigation, corrective action, and, where necessary, modification or withdrawal of the device.

Ongoing clinician review is also important to maintaining trust of both clinicians and patients, who need confidence that clinically consequential GenAI performance continues to be evaluated by people capable of assessing the clinical significance and reliability of its outputs, with clear manufacturer accountability for ensuring that this surveillance occurs and responding when it identifies a problem.

Professional medical organizations can play an important role in establishing best practices for this work, including appropriate clinician qualifications, sampling approaches, adjudication methods, and criteria for identifying clinically consequential performance changes. The AMA would welcome the opportunity to work with FDA and other stakeholders to develop these approaches and ensure they reflect the realities of clinical practice.

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

⁵ Babic B, Cohen IG, Stern AD, Li Y, Ouellet M. A general framework for governing marketed AI/ML medical devices. *NPJ Digit Med.* 2025;8(1):328. doi:10.1038/s41746-025-01717-9.

The characteristics of GenAI-enabled medical devices require a more coordinated and stringent approach to post market surveillance than reliance on individual manufacturer monitoring or traditional adverse-event reporting alone. FDA should consider establishing or adapting a national surveillance capability for GenAI-enabled medical devices that combine active surveillance, manufacturer-generated performance data, publicly available standardized reporting of adverse events and performance failures, and appropriately sampled clinical review.

Existing FDA surveillance approaches may provide useful models for developing such a capability, although they also require improvement. For GenAI-enabled devices, surveillance should be capable of identifying clinically inappropriate outputs, changing patterns of errors, performance degradation, and other failures that may not produce an immediately identifiable adverse event. Traditional monitoring and defined reportable events should be expanded to create a detailed record of instances of unexpected, anomalous, or questionable outputs, which should then be categorized by degree of potential clinical impact and studied for improvement. An expansion of reportable events would result in additional labor for HCPs and should be accompanied by sufficient labor and financial support.

FDA should retain responsibility for the national regulatory surveillance infrastructure, and manufacturers should remain responsible for monitoring and demonstrating the ongoing safety and effectiveness of their individual products. Monitoring programs should include simple mechanisms for reporting and timely feedback to reporters. A shared surveillance system should strengthen these responsibilities rather than diffuse them.

Physicians and professional medical organizations also have an important role. Clinical expertise will be necessary to define which performance signals are clinically meaningful, develop appropriate taxonomies for GenAI-related failures, establish methods for sample-based clinical review, and interpret whether observed changes have implications for patient care. Practicing physicians must remain at the forefront of decision-making regarding AI adoption and be actively engaged in reviewing and validating AI outputs to ensure clinical accuracy and safeguard patient safety. This expertise cannot be substituted and is essential for determining whether AI technologies meet the high standards required for healthcare.⁶ However, while physicians and professional medical societies can assist in defining clinical standards, the primary burden of continuous monitoring must lie with manufacturers, and participation in this ecosystem should not create a new uncompensated reporting or monitoring burden for individual physicians.

The AMA welcomes the opportunity to work with FDA, manufacturers, health systems, patients, and other stakeholders to develop the clinical components of such a surveillance framework, including definitions of clinically consequential AI failures and deficiencies, appropriate triggers for review, sampling and adjudication methods, and approaches for communicating emerging performance concerns to physicians.

Are there additional considerations that inform the premarket and post market 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?

⁶ American Medical Association. Statement of the American Medical Association to the U.S. Senate Committee on Health, Education, Labor and Pensions Regarding AI's Potential to Support Patients, Workers, Children, and Families. October 9, 2025. <https://searchlf.ama-assn.org/letter/documentDownload?uri=/unstructured/binary/letter/LETTERS/lfaip/2025-10-9-SFTR-re-Senate-HELP-AI-s-Potential-to-Support-Patients-Workers-Children-and-Families.pdf>.

Agentic GenAI-enabled systems involve extremely heightened risks compared to GenAI- and AI/ML-enabled medical devices. Just as GenAI-enabled medical devices carry more risk than traditional AI/ML-enabled devices, agentic GenAI-enabled systems pose greater risks than non-agentic GenAI-enabled devices given their heightened level of autonomy and independence without HCP oversight. Therefore, agentic GenAI-enabled systems require much higher levels of testing and scrutiny before being introduced into care settings. Additionally, we have concerns that the levels of HCP readiness, understanding, and evidence supporting use of these systems are inadequate for their deployment. Agentic GenAI-enabled systems are capable of autonomously planning and executing multi-step tasks, using external tools, and taking sequential actions, and the clinical consequences of their actions can be extreme. As such, these systems require special considerations and more in-depth conversations with stakeholders to determine appropriate regulatory requirements. The views of HCPs will be essential in these discussions, and we look forward to providing more input on approaches required for safe and effective deployment of agentic GenAI-enabled systems.

Thank you for considering our recommendations. The AMA remains committed to working with you to advance our shared goals and looks forward to continued collaboration to help recognize the potential of AI to improve the health of our nation. Please reach out to me directly at 312-464-5288 or John.Whyte@ama-assn.org if you have questions or need further information.

Sincerely,

A handwritten signature in black ink, appearing to read 'John Whyte', with a stylized, cursive script.

John Whyte, MD, MPH