AAFP
AMERICAN ACADEMY of FAMILY PHYSICIANS

October 15, 2026

The Honorable Kyle A. Diamantas, J.D.
Acting Commissioner
U.S. Food and Drug Administration
10903 New Hampshire Avenue
Silver Spring, MD 20993

RE: FDA Considerations for the Regulation of Generative AI-Enabled Medical Devices Request for Feedback; Docket ID: FDA-2026-N-7874

Dear Commissioner Diamantas:

On behalf of the American Academy of Family Physicians (AAFP), representing 124,500 family physicians and medical students nationwide, we appreciate the opportunity to provide feedback on the Food and Drug Administration's (FDA) discussion paper, Considerations for the Regulation of Generative AI-Enabled Medical Devices. The AAFP appreciates FDA's proactive effort to explore regulatory approaches that support innovation while ensuring that Generative AI (GenAI)-enabled medical devices are safe, effective, equitable, and trustworthy. As GenAI technologies become increasingly integrated into clinical workflows, patient interactions, and health system operations, **it is critical that regulatory frameworks remain focused on promoting patient safety, supporting evidence-based care, and preserving the patient-physician relationship.**

Family physicians are often the first point of contact within the health care system and provide comprehensive, continuous, and coordinated care across diverse patient populations and care settings. As a result, family medicine offers a unique perspective on both the opportunities and risks associated with GenAI-enabled technologies. Consistent with the AAFP's longstanding principles regarding artificial intelligence (AI), we believe these tools should be evaluated with the same rigor as other tools used in health care and should be designed to promote trustworthiness, privacy protection, and equity. We further believe that practicing physicians should be involved throughout the AI lifecycle, including development, validation, implementation, and postmarket evaluation to ensure that these technologies meaningfully improve patient care and clinical workflows.

The AAFP believes that clinical AI should augment, not replace, physician judgment and should preserve physician responsibility for clinical decision-making. For functions that initiate or materially influence clinical actions, the AAFP supports meaningful oversight and accountability by a licensed physician. The AAFP offers responses to questions most relevant to family medicine and primary care.

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

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

www.aafp.org

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

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.

The AAFP agrees that the FDA-identified additional dimensions of reversibility, downstream safeguards, time pressure, and traceability are appropriate risk modifiers to include.

The work of the Office of the National Coordinator for Health Information Technology (ONC) on the FDA Safety and Innovation Act in 2014 provided a list of 11 facets to consider for determining risk of health IT,[i] and we believe these are relevant for gauging AI risk. The facets identified are:

- purpose of the software product
- intended user(s)
- severity of potential injury
- likelihood of hazardous situation arising
- transparency of software operations and data
- ability to mitigate harmful conditions
- complexity of software (and maintenance)
- complexity of implementation and upgrades
- complexity of training and use
- use as part of a more comprehensive solution
- network connectivity, standards, and security

The AAFP recommends that the availability of meaningful physician oversight, patient health literacy, explainability and transparency, and the care setting also be included as risk modifiers. Meaningful physician oversight may reduce risk only when the physician has timely access to the information needed to evaluate an output; sufficient time and authority to independently review it; and the practical ability to accept, modify, reject, interrupt, or reverse a resulting clinical action before harm occurs, when clinically feasible. Oversight should not be treated as an effective safeguard when workflow design, time pressure, automation bias, inadequate transparency, or system permissions make physician review or intervention merely nominal.

Additionally, the AAFP believes organizational readiness and implementation capacity should also be considered when assessing risk. A safeguard that may be effective in a large, well-resourced health system may not function equivalently in an independent, rural, safety-net, or resource-constrained practice without adequate training, technical assistance, governance support, and implementation resources. FDA should evaluate whether a device's risk controls can be implemented effectively across the settings in which the manufacturer intends it to be used.

FDA could apply these modifiers after a device is initially placed within the two-axis framework, increasing the assigned risk level where traceability is limited, use is time-sensitive, the action is irreversible, or where meaningful physician review is absent. Effective safeguards may mitigate but should never be considered to eliminate the underlying risk of a high-consequence function.

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

For traditional Software as a Medical Device (SaMD), the output space is enumerable (e.g., a risk score has a range, an alert is or is not sent), and the manufacturer can characterize every output the device can produce. **A generative model cannot be characterized this way.** It produces outputs at runtime that vary with phrasing, conversational history, patient data, and stochastic sampling. A framework that asks manufacturers to classify the output is asking for something that does not exist as a fixed device attribute.

What manufacturers should be able to characterize, and what FDA can assess, are:

- the ceiling of directiveness the device is designed to reach;
- the controls that hold outputs at or below that ceiling (system prompting, output filtering, refusal behavior, template constraints, retrieval grounding); and
- the measured distribution of realized directiveness across representative queries and users.

FDA has already stated that directiveness "may depend on the substance and context of the output, not solely on whether it uses words such as 'recommend,' 'should,' or 'consider.'" The construct that actually predicts harm is not linguistic form but the probability that a user acts on the output without independent verification, which can be called effective directiveness.

For example, the same sentence "in situations similar to this, clinicians often increase the lisinopril dosage" can be interpreted as:

- a prompt for further thought to a hypertension specialist with time to think;
- a near-instruction to a family physician at minute 12 of a 15-minute visit with three other problems on the agenda and 40 messages in the inbox; or
- an instruction to a patient with no clinical training.

Risk may increase as outputs become more specific, urgent, tailored to an individual patient, or framed in language that implicitly encourages a particular action. The wording and the degree of confidence expressed also influence user behavior and should be weighed when assessing risk. Developers should clearly communicate intended use, limitations, performance characteristics, and validated uncertainty measures so physicians and patients can appropriately interpret outputs.

To improve regulatory predictability, the AAFP recommends FDA provide examples showing how these characteristics move an output along the continuum from general information to individualized instruction. FDA should also clarify that disclaimers alone do not convert an otherwise action-directing output into a non-directive function.

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 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?

The AAFP strongly supports efforts to expand patient access to health information and tools that facilitate patient engagement. Patient-facing GenAI-enabled functions may present different or elevated risks when their safe use depends on specialized clinical knowledge, when outputs are personalized or action-directing, or when an incorrect output could delay appropriate care. Risk assessment should focus on the design and consequences of the function rather than presume that patients are inherently unable to understand or act on health information.

We believe that "patient-facing" is too high-level a term to adequately represent the wide spectrum of scenarios that will have varying levels of risk. One critical factor is the degree to which the patient has a usual source of care and how that source of care has visibility and engagement with the "patient-facing" AI. For example, here are three scenarios that have low, moderate, and high risk, respectively:

Configuration	Recourse available	Risk profile
GenAI in a patient portal. Interaction visible to the patient's family physician with an established continuity relationship; appointment available.	Error is likely caught; patient has somewhere to take the output.	Low
GenAI in a health system application. No clinician sees the exchange, but the patient has an assigned physician.	Error caught only if the patient independently raises it.	Moderate
Standalone consumer GenAI. No clinician will ever know the conversation occurred; patient may have no usual source of care.	No correction pathway exists.	High

These applications differ by an order of magnitude in risk and are indistinguishable under a patient/health care provider (HCP) binary. Meanwhile, an HCP-facing output delivered to an AI company physician or locum covering an unfamiliar panel with no access to the patient's history will have less recourse than the first row above.

We believe manufacturers must include safeguards to help ensure the lowest risk to patients. In addition to integration with a patient's usual source of care, appropriate safeguards may include plain-language communication, disclosure of uncertainty, escalation pathways when high-risk symptoms are identified, and encouragement to engage with a clinician when appropriate. Patient-facing functions should also be designed and evaluated with varying levels of health and digital literacy in mind, with information presented at an appropriate reading level and available in the patient's preferred language when feasible. Patients should receive clear information about a device's intended use; limitations; and how information they provide will be collected, used, stored, or shared. Where additional assistance is needed, educational resources or other support should be available to help patients understand and use the technology with confidence. Such safeguards should empower patients without creating unnecessary barriers to access or reinforcing overprotective approaches to care. **Patient-facing tools should clearly disclose when information is AI-generated and distinguish general informational content from individualized medical advice.**

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?

Family physicians are trained to provide comprehensive, continuous care across the lifespan, and they routinely manage conditions that fall within multiple specialty domains. The AAFP believes GenAI-enabled tools have significant potential to extend specialty expertise into primary care settings, particularly in rural and underserved communities where specialty access is often limited or unavailable.

The AAFP is concerned that framing this issue as a distinction between generalist and specialist physicians may not accurately reflect the factors that influence risk. The relevant risk variable is not a deficit of knowledge on the part of generalist physicians, but a **mismatch between the competencies a device assumes of its user and the competencies its user actually possesses** — a mismatch that runs in both directions. A device calibrated for subspecialist use that presents as authoritative to a generalist is one failure mode; a device that oversimplifies for a physician who has substantial expertise in an area is another, and it produces the disengagement and alert fatigue that undermine safe use. A subspecialist receiving output outside their own field occupies precisely the position FDA describes, frequently with less longitudinal knowledge of the patient than the family physician with an established relationship would have.

Risk increases where a user has limited practical means of verifying an output. That condition correlates with access rather than with training: a family physician practicing in a county with no dermatologist within a hundred miles has materially less recourse than a clinician who can obtain an informal consultation the same day. **We therefore recommend that FDA require developers to define intended users by demonstrated competency and realistic conditions of use, rather than by specialty designation alone, and to validate device performance against those users' actual capabilities, practice settings, and available consultation resources**. Validation conducted in academic medical centers and large integrated systems does not establish performance in the ambulatory primary care settings where much of this technology will be used. The AAFP raised this point in our December 2025 response to FDA's request for information on real-world device performance, and it applies with equal force here.

We reject the false distinction that "generalist versus specialist" sets, as the boundaries between generalist and specialist practice are contested and heavily

overlapping. Management of diabetes, heart failure, depression, and common dermatologic and musculoskeletal conditions are simultaneously core family medicine practice areas and recognized specialty practice. We urge FDA to state explicitly that device labeling regarding intended users is not a determination of scope of practice and to avoid constructing a taxonomy of specialty-restricted clinical content. Defining intended users by competency and conditions of use rather than by specialty designation avoids this problem while still addressing the risk FDA has identified.

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?

The AAFP agrees that risk assessment for multi-turn conversational GenAI-enabled devices should consider the full conversational trajectory rather than individual outputs in isolation. A device that initially provides general information may become increasingly personalized, specific, or action-directing as additional context is provided. Risk may therefore arise from the cumulative effect of a conversation, even when no single response appears high-risk, and we recommend FDA evaluate whether device behavior remains consistent across longer interactions and whether clinically significant recommendations emerge only through the cumulative effect of multiple exchanges. These longer conversations can experience (1) user persistence, where a user asks for the same thing multiple times and the AI drifts toward compliance due to LLMs' tendency toward pleasing the user; (2) content accumulation, as more information that is patient-specific is added; and (3) stubbornness of AI LLMs, as current models are more likely to defend an earlier statement even if it is wrong than to change or revise it.

FDA should evaluate foreseeable conversational pathways, including those that result in individualized recommendations, triage advice, treatment suggestions, or other actions that could materially influence care. Factors such as personalization, expressed confidence, persistence of context, and the degree to which outputs direct a user toward a specific action should be considered in the assessment.

Intended use should reflect the device's reasonably foreseeable conversational behavior and demonstrated capabilities, not solely its labeling or individual response disclaimers. If a device can reasonably be expected to provide action-directing information through the course of a conversation, risk assessment should account for that function. Manufacturers should demonstrate performance and safety across

representative multi-turn interactions, including scenarios that test the boundaries of the device's intended scope.

Regarding physician oversight, where conversational outputs initiate or materially influence clinical actions, the AAFP believes oversight must remain meaningful. A physician presented only with a device's final recommendation cannot assess whether it emerged from sound reasoning, from an early error carried forward, or from user pressure over successive turns. **Meaningful oversight of a conversational device therefore requires access to the entire conversation, not merely to its conclusion.** We recommend FDA treat reviewability of the full exchange as a design requirement wherever a device's conversational output is intended to inform clinical action.

6. For GenAI-enabled care escalation functions, CDRH is considering how the evaluation may account for both under-escalation and over-escalation. How could manufacturers characterize and weigh these two directions of error, given that they may not be commensurable and that acceptable trade-offs may vary by clinical context?

The AAFP supports FDA's recognition that both under-escalation and over-escalation may cause harm. Failure to direct a patient to appropriate care can delay diagnosis and treatment, while unnecessary escalation can increase costs, burden emergency departments, contribute to patient anxiety, and erode trust. For high-acuity symptoms, the framework may appropriately place greater weight on avoiding under-escalation while continuing to monitor harms associated with excessive over-escalation. **Manufacturers should report these error types separately rather than combining them into a single measure, with prespecified acceptance thresholds based on the severity, reversibility, and likelihood of harm in the intended clinical context**.

However, we urge FDA to recognize that separate reporting alone will not produce a balanced picture, because the two error directions are not equally detectable. Over-escalation surfaces immediately and visibly; for example, the patient presents to an emergency department, testing is performed, the cost is incurred, and the error is readily attributed to the device. Under-escalation surfaces weeks or months later as a delayed presentation that no reporting system will connect to a triage interaction that preceded it, and that the patient does not know occurred. A manufacturer acting in good faith will therefore characterize over-escalation well and under-escalation poorly. The resulting data will give the appearance of strong performance, because the errors of greatest concern never enter the dataset. Additionally, the costs of over-escalation fall largely on emergency departments and primary care practices rather than on manufacturers. We encourage

FDA to not assume that manufacturer incentives are neutral between the two directions of error.

Under-escalation can be identified only prospectively; for example, by determining whether patients advised to monitor at home subsequently presented for urgent or emergency care, were admitted, or received a serious diagnosis within a defined follow-up period. While the AAFP recognizes that establishing such linkages is demanding and time-intensive, we do not believe there is an alternative, as complaint-driven and spontaneous reporting are structurally incapable of detecting the failure mode FDA identifies as a top concern. The AAFP recommends FDA require outcome-linked surveillance with a specified follow-up window for escalation functions, given that the interoperability infrastructure necessary to support such surveillance exists, although its availability, implementation, and maturity remain uneven across health care settings. We recommend FDA consider and account for the governance, technical assistance, and implementation support necessary to ensure this infrastructure can be used reliably for surveillance purposes.

Escalation performance is also dominated by prevalence in ways that directly impact primary care. An escalation function validated in an emergency department population, where the prevalence of serious pathology may be several percent, will behave very differently in an ambulatory primary care panel where it is a fraction of a percent. Positive predictive value falls sharply as prevalence declines even when sensitivity and specificity are unchanged, so a tool that appears well calibrated in its validation setting may generate many false escalations for each true one in family medicine, while the manufacturer's reported performance figures remain accurate. We recommend FDA require that escalation performance be reported together with the prevalence of the validation population and that manufacturers report expected performance at prevalences representative of the settings in which the device will actually be deployed. We wish to reiterate that validation conducted in academic medical centers and large integrated systems does not adequately establish performance in ambulatory primary care.

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

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?

Overall, the AAFP supports FDA's proposed competency-based approach combining device benchmarking and clinical confirmation. However, we caution against the use of the term "competency" to characterize the performance of GenAI-enabled

devices. In medical education, competency generally describes a physician's demonstrated ability to apply knowledge and clinical reasoning across a range of circumstances, while performance on a model benchmark does not necessarily demonstrate the same ability to generalize beyond the conditions on which the device was evaluated. A GenAI-enabled device may perform well on curated clinical questions but perform differently when presented with novel or out-of-distribution cases. **The AAFP encourages FDA to use terminology that clearly distinguishes benchmark performance from demonstrated clinical performance and to ensure that benchmarking does not substitute for robust clinical confirmation**.

Given the open-ended nature of many GenAI-enabled systems, traditional testing methods alone may not adequately characterize performance. The AAFP recommends that FDA maintain a risk-based approach and meaningfully engage practicing physicians throughout development, validation, implementation, and ongoing evaluation. When a device influences clinical care, evaluation should confirm that it augments rather than displaces clinical judgment and that an identifiable party remains accountable for its clinical effect. For GenAI-enabled devices integrated into clinical workflows, a licensed physician should retain responsibility and authority for individual clinical decisions while manufacturers remain accountable for a device's safety, performance, surveillance, and when necessary, remediation.

Clinical confirmation should assess not only technical performance and safety but also the effects of GenAI-enabled devices on care delivery. In primary care, relevant considerations include care coordination, patient understanding, workflow integration, physician workload and cognitive burden, and whether the technology supports meaningful patient-physician interactions. A device that performs accurately on individual tasks may nevertheless create unintended harms if its implementation fragments care, increases administrative burden, undermines clinical judgment, or disrupts clinical workflows.

8. 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 amount and rigor of evidence required should be proportional to the independence of the device's activity, the consequences of an incorrect output, and the device's potential impact on patient care. Higher-risk applications, including those that influence diagnosis, treatment, prescribing decisions, or triage should require more robust evidence than functions demonstrated to have limited clinical consequences, regardless of whether they are characterized as administrative or informational. **We emphasize that functions labeled administrative frequently carry**

clinical consequence – both a scheduling function that deprioritizes a symptomatic patient and a documentation function that shapes what enters the record can impact patient care, whether or not it is described as clinical.

We encourage FDA not to collapse these two axes into a single tier, as the nature of the evidence required should reflect the distinct risks presented by each axis. A function with limited independence but severe consequences and a function with high independence but modest consequences may occupy similar positions on a combined scale while requiring different types of evidence, rather than different amounts. The severity of potential consequences should inform the evidence necessary to demonstrate accuracy, characterize potential failures, and assess the consequences of an incorrect output. The degree of activity independence should inform the evidence necessary to demonstrate appropriate safeguards, human factors considerations, reviewability, and opportunities for meaningful human intervention. We therefore encourage FDA to evaluate these considerations separately rather than combining them into a single measure of evidentiary burden.

The AAFP recommends that FDA consider a third factor that the two axes do not currently capture, which is the breadth and variability of outputs a device may generate. A narrowly scoped generative function and a general-purpose conversational device may present similar levels of risk under the two-axis framework while requiring substantially different evidence to adequately characterize their performance and potential risks. Evidence requirements should account for the range and variability of a device's potential outputs and be transparent and predictable while maintaining flexibility for future innovations.

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 benchmarking structure is a useful foundation, but it is unlikely to provide sufficient evidence of safety and effectiveness on its own. FDA appropriately recognizes that structured benchmarking cannot fully capture how GenAI-enabled devices interact with users, workflows, patient populations, and deployment environments, underscoring the need for clinical confirmation.

The AAFP recommends FDA address an important methodological limitation associated with the use of publicly available benchmarks, given that published benchmark materials may be included in the training data used to develop GenAI

models. As a result, strong performance on a publicly available benchmark may reflect prior exposure to the benchmark rather than the device's ability to perform the underlying task, making it difficult to determine what the reported score demonstrates. To ensure benchmarking provides meaningful evidence of device performance, the AAFP recommends that FDA require evaluation using appropriate materials that are not available to developers, are updated on a defined schedule, and have documented provenance to minimize the likelihood they were included in training data.

The AAFP believes several important elements should be strengthened or made more explicit, including performance across multi-turn conversations, workflow usability, automation bias and user over-reliance, traceability to supporting evidence or source materials, and users' ability to recognize and appropriately respond to incorrect outputs. These factors can substantially affect real-world safety but may not be fully reflected in the proposed categories. We recommend FDA further explore the use of externally developed consensus standards and qualified third-party evaluation programs where appropriate, through a transparent, multi-stakeholder process that promotes consistency, independence, and scientific validity.

The AAFP recommends FDA clearly define generalizability to include performance across different practice settings, patient populations and disease prevalence, languages, and changes in clinical evidence over time. Evaluation should also account for changes in clinical guidelines and standards of care to ensure that benchmark results continue to reflect current evidence-based practice.

Regarding external standards, the AAFP would support FDA's engagement with existing frameworks including the National Institute of Standards and Technology's AI Risk Management Framework and relevant consensus work by health IT standards bodies through a transparent, multi-stakeholder process.

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?

The AAFP supports FDA's recognition that prospective clinical trials may not be necessary in every circumstance. Sponsors should select and justify the clinical

confirmation method based on the device's risk, as determined by a risk-based framework like we've proposed throughout this response, including intended use and users, clinical consequences, degree of autonomy, deployment setting, and variability of anticipated inputs. Confirmation should use inputs representative of the populations, workflows, and conditions expected in actual use.

Regarding the anticipated distribution of real-world inputs, we note that for generative devices this distribution cannot be enumerated in advance and that harm concentrates in its tail rather than its center. Evaluation should include atypical clinical presentations, incomplete or unclear user inputs, non-English language inputs, and other reasonably foreseeable scenarios in which device performance may differ from expected performance. We recommend FDA require sponsors to describe how these scenarios were incorporated into clinical confirmation and demonstrate that evaluation was not limited to the most common or expected inputs.

Retrospective analyses, shadow deployments, or physician review may provide sufficient clinical confirmation for some lower-risk functions. However, the AAFP encourages FDA to recognize the limitations of retrospective confirmation when a device's primary risks involve errors of omission. Retrospective review may identify incorrect outputs that are documented in the clinical record but may be less effective at identifying instances in which a device fails to identify a clinically significant finding or recommend appropriate escalation. When missed findings or failures to escalate could result in significant patient harm, retrospective analysis alone should not be considered sufficient clinical confirmation.

Prospective clinical evaluation should be required when a device initiates or materially influences diagnosis, treatment, prescribing, emergency-care direction, or another action with a meaningful risk of serious or irreversible harm. Evidence of performance, however, should not by itself justify autonomous exercise of clinical judgment without meaningful physician oversight. Real-world evaluation should complement, rather than substitute for, evidence appropriate to the device's premarket risk.

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 may be useful for supplementing evaluation of rare conditions or scenarios that are difficult to observe at sufficient scale. Synthetic augmentation is

better suited to rare but well-characterized presentations, to adversarial and stress testing, and to privacy-preserving augmentation of populations already well represented in the source data. Safeguards should include comparison against independent real-world datasets, subgroup-specific performance reporting, documentation of the generation process and source-data limitations, and evaluation by experts independent of the model developer. **Synthetic data generated by a model from the same class as the device should not serve as the sole basis for evaluating the performance gaps it is intended to detect, particularly for populations that have historically experienced disparities in access, outcomes, or representation in training data.**

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 and acceptance criteria should reflect the device's intended use, users, deployment setting, and clinical consequences. Qualified physician panels should include expertise in both the clinical domain and intended users. Specialist input may be necessary when safe performance depends on specialty-level knowledge but should not automatically be used for a tool intended to support a generalist practice.

The applicable comparator should generally reflect an evidence-based, clinically appropriate standard of care rather than median clinician performance, which may reflect unsafe or outdated practices. When physician review or intervention is part of the authorized use, human-AI team performance should be the primary basis for evaluation because it best reflects real-world deployment. However, device-only performance should also be assessed to identify intrinsic failure modes, characterize automation-bias risks, and evaluate the potential consequences of foreseeable over-reliance or reduced human oversight.

When human-AI team performance is used as the basis for evaluation, the AAFP recommends that participating physicians be representative of the device's intended users and that evaluation reflect realistic conditions of use. Studies that rely primarily on clinicians with specialized training or experience using the device may not adequately reflect performance in routine clinical practice. **Evaluation should include physicians with varying levels of experience with the technology and reflect realistic clinical conditions, including time constraints, workload, and workflow demands**.

As discussed further in our response to Question 15, the AAFP recommends FDA use these comparators for distinct but complementary purposes. An evidence-based standard of care should inform whether device performance is clinically appropriate, while the care pathway likely to occur without the device can help determine whether the device improves upon existing care in its intended setting. Performance that exceeds the likely alternative care pathway should not compensate for failure to meet an appropriate clinical standard.

15. 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?

Evaluation should compare performance with the care pathway most likely to occur for the intended population and setting in the absence of the device, such as unaided clinical judgment, delayed specialist consultation, or lack of access to specialty expertise. Sponsors should justify that pathway using clinical workflow data, published evidence, or real-world utilization data and should avoid selecting an artificially weak comparator that exaggerates the device's apparent benefit. The selected comparator should reflect actual patterns of care and access for the population and setting in which the device is intended to be used, including circumstances in which patients may face delayed or limited access to specialty care. Where the expected alternative course of care differs meaningfully across populations or care settings, evaluation should account for that variation rather than relying on a single comparator that may not accurately represent real-world care. Comparisons should also consider clinically meaningful patient outcomes and potential harms, rather than focusing solely on whether the device performs a particular task more accurately or efficiently than the alternative.

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?

Independent third parties may be particularly valuable for validating benchmark design, evaluating subgroup performance, adjudicating open-ended clinical outputs,

and confirming sponsor-reported results. Participating organizations should demonstrate relevant technical and clinical expertise and disclose financial, contractual, governance, or data relationships with the manufacturer and any foundation-model developer. Where third-party evaluators or benchmarking organizations are utilized, FDA should support inclusion of practicing physicians and other end-users in benchmark development, test design, and evaluation activities to help ensure assessments reflect real-world clinical practice and patient-care needs. We also encourage FDA to name specialty societies specifically among the parties whose participation it should seek. Societies maintain clinical guidelines, convene practicing clinicians across diverse settings, and can contribute domain content and reviewers without the commercial relationships that complicate other forms of third-party participation. The AAFP would welcome the opportunity to contribute to family medicine and primary care domains.

Third-party review may be impractical when proprietary-system access or sensitive clinical data cannot be provided safely. To avoid limiting competition, FDA should establish common qualification criteria and transparent methods while avoiding exclusive accreditation arrangements, excessive fees, or requirements that disproportionately disadvantage smaller developers. Appropriate review and appeal processes should also be available.

Section VI: Postmarket Monitoring for GenAI-Enabled Devices

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?

The AAFP does not believe increased postmarket monitoring should routinely substitute for adequate premarket evidence in higher-risk clinical applications. Devices that materially influence diagnosis, treatment, prescribing, or emergency care decisions must be supported by robust premarket evidence prior to widespread deployment. Greater reliance on postmarket monitoring may be appropriate for lower-risk GenAI-enabled devices when there is a well-defined intended use; limited potential for patient harm; and a comprehensive monitoring program that includes predefined performance thresholds, active surveillance, adverse event reporting, performance degradation monitoring, and timely corrective action processes.

To justify reduced premarket evidence, manufacturers should demonstrate the ability to detect clinically meaningful performance changes and implement effective

mitigation measures before patient harm becomes widespread. Monitoring programs must include ongoing evaluation across relevant patient populations and deployment settings, with clear accountability for responding to identified risks.

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?

The AAFP supports periodic re-benchmarking, physician review, and monitoring for performance degradation. FDA should also consider additional triggers for reassessment, including:

- Clinically significant performance drift;
- Changes in patient population or practice settings;
- Meaningful model or workflow updates;
- High or low override rates;
- Safety events, near misses, or clinician-reported concerns; and
- New clinical guidelines or standards of care.

With that said, we believe the override-rate signal trigger needs more specificity. A high override rate may indicate poor device performance, poor workflow integration, or clinician distrust, with each warranting different responses. A very low override rate warrants equal attention, since it may reflect either excellent performance or uncritical acceptance, and these are not distinguishable from the rate alone. We recommend FDA treat unexpected values in either direction as a trigger for investigation rather than monitoring only the high end.

The routine reassessment cadence should be risk-based and informed by the device's clinical consequences, degree of autonomy, observed rate of change, volume of use, and stability of the underlying model and deployment environment. Event-triggered reassessment must occur promptly when predefined thresholds are met, rather than waiting for the next scheduled review.

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?

The AAFP believes machine-based supervisory agents may help facilitate postmarket monitoring activities such as performance degradation, trend detection, audit-log review, and identification of potential safety signals. Such tools may improve the scalability and timeliness of postmarket surveillance programs. **However, supervisory agents should support, not replace, human oversight and clinical judgment.** Findings generated through automated monitoring should be reviewed and approved by qualified individuals before they are relied upon for safety determinations, regulatory reporting, corrective actions, or other consequential decisions. Manufacturers should remain responsible for monitoring device performance, investigating identified risks, and ensuring continued safety and effectiveness throughout the product lifecycle.

FDA should evaluate these supervisory agents as components of the monitoring system and require evidence that they can reliably detect clinically meaningful performance changes and appropriately escalate concerns for human review. Evaluation should address sensitivity, false positive and false negative rates, performance drift, transparency, auditability, and performance across diverse populations and deployment settings. The role, limitations, and oversight processes associated with a supervisory agent should be clearly defined and documented. **While automated surveillance may strengthen postmarket monitoring, it should not eliminate meaningful human review or responsibility for patient safety.**

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?

Physicians, health care organizations, professional societies, standards development organizations, and researchers can help identify safety and usability concerns, report performance issues, and inform best practices. These activities should complement rather than replace manufacturer responsibilities to monitor performance, investigate risks, implement corrective actions, and maintain safety and effectiveness throughout the product lifecycle.

We recommend postmarket monitoring programs include simple reporting pathways, timely feedback, and adequate technical and financial support for participating practices. Such programs should not rely on uncompensated physician labor, nor should they transfer investigation or remediation responsibilities away from manufacturers and onto frontline physicians.

The AAFP encourages FDA to consider potential legal and liability concerns that may discourage physicians from reporting device-related safety events. Physicians may be

reluctant to report events involving their own patients if they are concerned that such reports could be used in subsequent litigation. These concerns may present particular challenges for physicians in small and independent practices that have limited access to institutional legal or risk-management resources. We encourage FDA to explore appropriate protections for good-faith physician reporting, consistent with existing patient safety reporting frameworks, to support robust postmarket surveillance.

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?

The AAFP reiterates the position stated in our December 2025 response to the FDA's request for information on real-world device performance: physicians must be notified when device behavior changes materially under an approved predetermined change control (PCCP) plan. A physician relying on a device in clinical care cannot exercise meaningful judgment about that reliance if the device's behavior has changed and they have not been notified. This concern is heightened for generative devices, where a modification may alter the substance, specificity, or directiveness of outputs without any change to the device's labeling, indications, or user interface. We recommend FDA establish notification expectations as a condition of PCCP approval.

Risk-based reevaluation should accompany changes that could materially affect clinical performance, safety, usability, patient outcomes, intended use, autonomy, clinical recommendations, or performance in specific populations. Administrative or technical changes that do not alter intended use, clinical workflow, access to care, system permissions, data sources, output behavior, or performance may be managed through established quality system processes. An "administrative" designation alone should not determine the level of review. A PCCP may be appropriate for bounded, foreseeable changes when the modification protocol, validation methods, impact assessment, monitoring expectations, and acceptance criteria can be specified in advance. **Changes that materially expand intended use or increase the severity of potential harm should be required to receive additional FDA review**.

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?

Manufacturers should have sufficient contractual and technical visibility into material model changes to evaluate how those changes to an underlying third-party foundation model affect device performance, safety, and intended use. A manufacturer should not deploy or continue using an updated foundation model when it lacks sufficient notice, documentation, access, or technical controls to determine whether the change affects the authorized device.

Relevant mechanisms may include contractual requirements for advance notice of material changes; version control and change logs; access to information necessary for impact assessment; technical monitoring for unexpected behavioral changes; the ability to suspend or roll back deployment; and predefined revalidation procedures.

When foreseeable model changes are addressed through a PCCP, the plan should identify the permitted scope of change, testing requirements, acceptance criteria, monitoring obligations, and conditions requiring FDA review. Manufacturers should also disclose how data transmitted to a third-party model provider may be collected, retained, used, or shared, and they should protect identifiable patient information consistent with applicable privacy and consent requirements.

The AAFP encourages FDA to consider whether these safeguards are reasonably available to manufacturers of different sizes and with varying levels of negotiating leverage. Smaller developers, including those developing tools for independent and primary care practices, may have less ability to negotiate contractual provisions regarding advance notice of model changes, change documentation, or other protections from third-party foundation model providers. FDA should ensure that its regulatory approach does not depend solely on a manufacturer's ability to negotiate such terms and should provide clear expectations for alternative technical safeguards when appropriate.

We encourage FDA to consider whether certain baseline obligations should be attached to foundation model providers whose models are incorporated into authorized devices, rather than depending on the device manufacturer's negotiating position. Where such terms cannot be obtained, FDA should specify what technical monitoring would be accepted as a substitute, so that inability to negotiate favorable contractual terms does not by itself foreclose participation.

Section VII: Other Topics

26. Are there additional considerations that inform the premarket and postmarket evaluation of agentic GenAI-enabled devices, beyond those applicable to nonagentic 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 GenAI-enabled devices warrant heightened scrutiny because they may autonomously execute multistep actions, use external tools or data, and propagate an error across an entire workflow before a person has an opportunity to intervene. The degree of permitted autonomy should therefore be proportionate to the function's clinical risk and the reversibility of its actions. Functions that initiate or materially influence high-consequence clinical actions should require meaningful oversight and accountability by a licensed physician, including a practical opportunity to review, override, or reverse an action before harm occurs, when clinically feasible.

Evaluation should assess the safety of the complete action sequence, not merely the accuracy of individual outputs. Agentic systems should demonstrate an ability to recognize when a task exceeds their authorized scope, respond appropriately to tool failures or conflicting information, and reliably escalate issues requiring human review. Testing should also assess foreseeable combinations of individually minor errors that could create significant cumulative harm.

The AAFP emphasizes that the opportunity for human review must be meaningful in practice in order to serve as an effective safeguard. Review processes that require physicians to evaluate numerous agent-generated actions under significant time or workload constraints may limit the effectiveness of human oversight. As shared elsewhere throughout this response, we believe that when human review is relied upon as a risk-control measure, manufacturers should be required to demonstrate that physicians have sufficient time, information, and ability to meaningfully review, modify, or reject device actions under realistic conditions of use.

The AAFP recommends that agentic GenAI-enabled devices maintain a traceable record of actions performed by the device, including the sequence of actions and the information relied upon at each step. This information is important to support meaningful physician review, facilitate investigation of safety events, and identify where errors may have occurred within a multi-step process. **We strongly urge FDA to consider such traceability as a foundational, essential safety and accountability requirement for agentic functions, particularly when a device initiates or**

materially influences clinical actions. For systems that invoke other models, agents, tools, or external services, traceability should extend across the full chain of activity. Manufacturers should identify which component performed each action, what data were exchanged, how permissions were inherited or expanded, and how errors or unauthorized disclosures can be contained across interconnected systems.

Conclusion

Thank you for the opportunity to comment on this important discussion paper and its potential impacts on primary care in America. The AAFP supports FDA's efforts to develop a risk-based, total product lifecycle approach to the oversight of GenAI-enabled medical devices. While these technologies hold significant promise to improve patient access and clinical efficiency, their successful integration into health care requires robust evaluation, ongoing monitoring, transparency, and meaningful clinical oversight. **We believe regulatory expectations should account not only for the accuracy of individual outputs, but also for autonomy, actionability, permissions, cumulative effects, deployment context, and the device's capacity to influence or execute clinical actions.** Regulatory approaches should remain focused on ensuring that innovation advances patient care without creating new risks, exacerbating disparities, or undermining trust in the health care system.

We encourage FDA to continue engaging stakeholders as these policies evolve and stand ready to serve as a resource and partner in developing evidence-based frameworks that support the responsible deployment of GenAI-enabled medical devices. For more information or questions, please contact Mandi Neff, Senior Strategist, Regulatory and Policy, at mneff2@aafp.org.

Sincerely,

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

[i] Slight SP, Bates DW. A risk-based regulatory framework for health IT: Recommendations of the FDASIA working group. J Am Med Inform Assoc. 2014;21(e2):e181-e184. doi:10.1136/amiajnl-2014-002638.