



October 19, 2026

Dockets Management Staff (HFA-305)
U.S. Food and Drug Administration
5630 Fishers Lane, Rm. 1061
Rockville, MD 20852

**RE: Considerations for the Regulation of Generative AI-Enabled Medical Devices:
Discussion Paper and Request for Feedback [FDA-2026-N-7874]**

Submitted electronically via regulations.gov

Dear Acting Commissioner Diamantas:

The National Health Council (NHC) appreciates the opportunity to provide comments on the Food and Drug Administration's (FDA's) discussion paper, Considerations for the Regulation of Generative AI-Enabled Medical Devices. The discussion paper addresses a rapidly evolving area of medical product development and appropriately recognizes both the potential for GenAI-enabled devices to improve care and the need to address risks associated with open-ended outputs, evolving models, third-party foundation models, and increasing levels of autonomy. The NHC supports FDA's effort to seek early stakeholder input and to consider a risk-based, total product lifecycle approach that can support timely access to safe and effective devices.¹

The NHC unites nearly 200 national organizations—including leading patient groups, research institutions, providers, caregivers, and businesses across the health care sector—to drive patient-centered health policy. Representing 200+ million Americans with chronic diseases and disabilities, the NHC strengthens its members' collective influence to expand access to quality, affordable, and equitable health care. The NHC fosters collaboration to shape policies that reflect the needs of patients.

Across its work on artificial intelligence (AI) in health care, the NHC has emphasized that innovation should improve patient care while maintaining safety, effectiveness, transparency, accountability, privacy, and meaningful human involvement. In testimony to Congress, the NHC emphasized that AI should support and supplement—not supplant—human decision-making and patient preferences. The testimony also highlighted the importance of pre-deployment testing, meaningful oversight, understandable information, and continuous feedback are to responsible AI use. The NHC's *Principles on Health AI* identify patient health benefit, patient-centeredness,

¹ U.S. Food and Drug Administration, Considerations for the Regulation of Generative AI-Enabled Medical Devices: Discussion Paper and Request for Feedback (August 2026), 1–3, <https://www.fda.gov/media/189158/download>.

education and empowerment, equity and accessibility, trustworthiness, comprehensive transparency, bias mitigation, privacy and security, human collaboration, continuous improvement, safety and effectiveness, accountability, and ethical considerations as foundational expectations for health AI.^{2,3}

These principles are particularly relevant to GenAI-enabled medical devices. Many of the characteristics that may make GenAI useful—including flexibility, natural-language interaction, personalization, and multimodality—can also complicate efforts to evaluate performance and identify failures. The NHC supports a regulatory framework that is risk-proportionate, scientifically rigorous, predictable, and sufficiently flexible to accommodate technological change. As FDA develops that framework, the NHC encourages the Agency to incorporate patient experience and patient-facing consequences throughout risk assessment, premarket evaluation, change control, and postmarket monitoring.

Summary of Recommendations

- Use the proposed two-axis risk framework as a starting point while accounting for additional considerations, including reversibility, detectability of error, time sensitivity, traceability to source information, reliance and automation bias, user context, and the availability of meaningful human review.
- Evaluate patient-facing functions based on clinical context and the consequences of reliance rather than presuming higher risk solely because the intended user is not a health care professional. Relevant considerations include the directiveness and personalization of the output, the user's practical ability to recognize uncertainty or error, and the availability of appropriate safeguards.
- Support a competency-based premarket framework that combines risk-proportionate benchmarking with clinical confirmation and aligns the amount and type of evidence with intended use, real-world conditions, and the consequences of error.
- Encourage evaluation across clinically relevant populations and deployment contexts, including differences in health literacy, language, disability, age, geography, and other factors that can affect how patients interact with and interpret GenAI outputs. Synthetic data can be a useful supplement but should not routinely replace evidence from real patients and real-world settings.
- Incorporate patient input into communication, comprehension, usability, and escalation benchmarks, as well as judgments about meaningful performance and

² National Health Council, *Principles on Health AI*, March 25, 2025, <https://acrobat.adobe.com/id/urn:aaid:sc:va6c2:1e4c2cb5-5b01-4e65-b386-8eeefea9b6f5>.

³ National Health Council, "NHC Statement on Artificial Intelligence and Health Care: Promise and Pitfalls," statement for the record submitted to the Senate Committee on Finance, February 8, 2024, <https://nationalhealthcouncil.org/letters-comments/nhc-statement-on-artificial-intelligence-and-health-care-promise-and-pitfalls/>.

acceptable uncertainty. This input should help ensure that regulatory evaluation reflects how patients and caregivers understand and act on GenAI-generated information.

- Use postmarket monitoring as a central component of total product lifecycle oversight . Allow greater reliance on postmarket evidence to address premarket uncertainty only when the device's risk, reversibility, monitoring capability, and available intervention pathways make that approach appropriate.
- Encourage transparent, risk-based postmarket monitoring that can detect performance drift, subgroup disparities, recurrent failure modes, and changes in patient-facing behavior, with clear triggers for reassessment and mechanisms for patients and clinicians to report concerns.
- Maintain clear manufacturer accountability when responsibility is shared across foundation-model developers, health systems, clinicians, standards organizations, and other ecosystem participants. Contracts, audit logs, update notifications, version controls, and predetermined change control plans can help manage these dependencies while preserving clear responsibility for the regulated device.
- Explore voluntary Foundation Model Device Master Files as a potentially useful mechanism to improve regulatory efficiency and consistency, particularly when they provide standardized, current, and versioned information on model characteristics, limitations, safety controls, update practices, and subgroup performance. The device sponsor should retain responsibility for safety and effectiveness of the final deployed device.
- Use additional, risk-proportionate safeguards for agentic GenAI-enabled devices as autonomy and the consequences of action increase, including clearly defined permissions, meaningful human checkpoints before irreversible or high-consequence actions, auditable action logs, fail-safe behavior, rollback mechanisms, and evaluation of cumulative multi-step failure and cybersecurity risks.
- Maintain a technology- and architecture-neutral framework while allowing testing expectations to reflect architecture-specific failure modes for multimodal, vision-language, predictive, and other emerging models when those differences may affect safety, effectiveness, or patient understanding.

Risk Assessment Should Reflect Patient Context and Real-World Consequences

The NHC supports FDA's proposal to assess risk based on the core dimensions of device activity and the consequences of relying on an incorrect output. These dimensions appropriately distinguish among tools that provide information, tools that direct users toward a particular action, and tools that take action with limited or no human intervention. They also recognize that similar levels of autonomy may carry different risks depending on the clinical consequences of error. The NHC encourages

FDA to organize risk assessment around these dimensions while allowing additional factors to modify the resulting risk determination.

Additional considerations may include reversibility, detectability of error, time sensitivity, traceability, and the availability of downstream safeguards. An incorrect output that is readily recognizable, easily corrected, and unlikely to cause harm before a clinician can intervene presents a different risk profile from an output that appears authoritative, cannot be independently checked, and may prompt an irreversible action under time pressure. Traceability is particularly relevant for GenAI because a fluent response may convey more certainty than the underlying evidence supports. FDA's framework would benefit from considering whether the intended user can understand the basis and limitations of the output, recognize uncertainty, and obtain meaningful review before acting.

This approach is consistent with the NHC's earlier comments on FDA's AI-enabled device software function guidance, where we encouraged rigorous validation, clear disclosure of intended use and limitations, subgroup-specific performance assessment, and ongoing verification throughout the total product lifecycle. It also reflects concerns raised by the patient and medical technology communities regarding algorithmic bias, clinical transparency, distribution shift, cybersecurity, privacy, and the importance of appropriate human review.^{4,5,6}

Patient-Facing Functions Warrant Context-Specific, Patient-Centered Evaluation

The discussion paper appropriately considers whether patient-facing informational functions may present different risks from functions intended for clinicians. It also recognizes the potential benefits of improved access to information, patient empowerment, and engagement. The NHC appreciates this balanced framing. Patients vary widely in health literacy, disease expertise, lived experience, access to clinicians, and familiarity with digital tools, and many people living with chronic or rare disease develop substantial expertise in their own conditions. FDA's framework should therefore avoid treating patient-facing use as categorically higher risk based solely on the intended user's professional credentials.

At the same time, GenAI systems can generate highly personalized and persuasive language that patients may reasonably interpret as tailored medical advice. The NHC encourages FDA to evaluate how strongly an output directs a user toward a particular action, considering specificity, personalization, confidence, context, repetition across a conversation, and whether the model adapts its recommendation after receiving

⁴ Alzheimer's Association, "Technology Safety for Older Adults," accessed October 6, 2026, <https://www.alz.org/help-support/caregiving/safety/technology-safety-older-adults>.

⁵ Advanced Medical Technology Association (AdvaMed), The Role of Artificial Intelligence (AI) in Healthcare, November 11, 2024, <https://www.advamed.org/member-center/resource-library/the-role-of-artificial-intelligence-ai-in-healthcare/>.

⁶ National Health Council, "NHC Submits Comments on FDA Draft Guidance for AI/ML-Enabled Medical Devices," April 7, 2025, <https://nationalhealthcouncil.org/letters-comments/nhc-submits-comments-on-fda-draft-guidance-for-ai-ml-enabled-medical-devices/>.

additional patient information. A generic explanation of treatment options, for example, presents different considerations from a response that integrates an individual's symptoms and medications and recommends a specific course of action.

The relevant regulatory question is therefore whether the intended user has a realistic opportunity to understand the output's limitations, recognize uncertainty or error, and obtain appropriate review before consequential action occurs. Depending on intended use and risk, safeguards may include plain-language explanations of uncertainty, source traceability, explicit boundaries on intended use, escalation pathways, confirmation steps, and human review for higher-consequence recommendations. Consistent with the NHC's *Principles on Health AI*, these safeguards can protect patients while preserving the benefits of access to useful information.⁷

FDA would also benefit from distinguishing disclosure from consent and tailoring patient-facing information to the role the system plays in care. Not every use of a GenAI-enabled device will warrant the same type or depth of explanation. When a function materially shapes a diagnosis, treatment recommendation, triage decision, or other consequential interaction, however, patients should receive understandable information about what the tool is intended to do. That information should explain whether its output is advisory or action-directing, the role of any human reviewer, important limitations, and how to seek review if the output appears inconsistent with their experience. The NHC encourages FDA to consider comprehension and usability, rather than the presence of disclosure alone, when evaluating patient-facing communication.

These considerations also apply when a caregiver is the primary or frequent user of a device. In pediatric care, dementia, disability, serious illness, and other settings, family caregivers may help interpret information and support decision-making. Evaluation should reflect these real-world patterns of use, including whether information is understandable to the person interacting with the device, whether escalation instructions are actionable, and whether the system appropriately distinguishes the roles of patients and caregivers. A patient-centered approach can account for caregiver involvement without treating it as a substitute for patient autonomy where the patient can participate directly.

A similar approach is appropriate when considering differences between generalist and specialist clinicians. The NHC does not believe specialty status alone should determine risk. Risk may increase when safe use depends on knowledge the intended user cannot reasonably be expected to possess, when an output falls outside the user's usual scope of practice, or when specialist-level conclusions are presented without sufficient information about uncertainty, limitations, or the need for consultation. FDA should consider the knowledge reasonably available to the intended user, the clinical consequences of error, and whether the device provides adequate context and pathways for referral or consultation. A GenAI-enabled function that responsibly extends

⁷ NHC, *Principles on Health AI*. See Principle 2: Patient-Centered and Informed; Principle 3: Patient Education and Empowerment; Principle 5: Trustworthiness; Principle 6: Comprehensive Transparency; and Principle 9: Human Collaboration.

access to specialty knowledge may meaningfully benefit patients, particularly in settings where specialist access is limited.⁸

Multi-Turn Conversations and Care Escalation Should Be Evaluated Under Realistic Conditions

The NHC agrees that GenAI risk cannot always be assessed one interaction at a time. A conversational device may move from general education to personalized interpretation and then to action-directing guidance over the course of an exchange. Evaluation would therefore benefit from testing realistic conversations that include long interactions, contradictory information, changing symptoms, emotional distress, repeated requests for out-of-scope recommendations, and circumstances in which a user minimizes or exaggerates symptoms. The conversation as a whole should remain consistent with the device's intended use and safety parameters.

For care escalation functions, both under-escalation and over-escalation can harm patients. Failure to identify an emergency can delay necessary treatment, while systematic over-escalation can create anxiety, unnecessary testing, avoidable emergency department use, financial burden, and reduced trust. The NHC encourages FDA to allow these errors to be weighed differently depending on clinical context rather than relying on a single undifferentiated measure of performance.

Evaluation of escalation would also benefit from clinically interpretable information beyond a single sensitivity or specificity estimate. Sponsors could characterize the types of cases in which under- and over-escalation occur, the severity and reversibility of resulting harms, and whether errors are concentrated in particular populations, symptom presentations, or conversational patterns. Aggregate performance can obscure clinically important differences. FDA should therefore encourage reporting that helps reviewers understand relevant tradeoffs and identify where additional safeguards or evidence may be needed.

A Competency-Based Premarket Framework Can Support Flexible, Fit-for-Purpose Evaluation

The NHC supports FDA's exploration of a competency-based approach that combines device benchmarking with clinical confirmation. Exhaustive testing of every possible input and output is unlikely to be practical for open-ended GenAI systems. A structured set of competencies may provide a scalable way to evaluate safety behavior, clinical proficiency, communication, robustness, subgroup performance, and agentic capabilities. The approach is most useful when it applies to the final deployed configuration and when the rigor of evaluation is proportionate to intended use and risk.

Benchmarking alone, however, may not fully demonstrate how a device will perform in the settings where patients and clinicians use it. Public benchmarks may be affected by

⁸ American Diabetes Association, "New ADA Scientific Statement: Diabetes Technology in Primary Care," September 28, 2026, <https://diabetes.org/newsroom/press-releases/new-ada-scientific-statement-diabetes-technology-primary-care>.

contamination or saturation, and benchmark datasets may not reflect the workflows, language patterns, comorbidities, or uncommon but consequential scenarios encountered in practice. The NHC therefore supports FDA's inclusion of clinical confirmation and encourages the Agency to treat benchmarking and clinical confirmation as complementary sources of evidence. Benchmarking can efficiently test broad capabilities and known failure modes, while clinically representative evidence can help establish whether those capabilities translate into safe and effective real-world performance.

The NHC also encourages FDA clarify the characteristics of fit-for-purpose benchmarks used to support regulatory decisions. Relevant considerations include whether a benchmark measures a capability meaningfully connected to the intended clinical use; reflects the difficulty, input distribution, and failure modes expected in practice; uses prespecified scoring and acceptance criteria; and produces reproducible results. Sequestered or independently maintained test sets may be useful where they reduce contamination or overfitting. Sponsor-developed benchmarks may also be appropriate when no suitable external benchmark exists, provided that sponsors explain the benchmarks' development methods, data provenance, representativeness, and relationship to clinically relevant performance.

This position builds on the NHC's comments responding to an NIH RFI, in which the NHC supported external validation on diverse, representative datasets; standardized performance benchmarks; simulation environments for edge cases; third-party auditing tools; and independent evidence of safety, effectiveness, and equity before clinical deployment of federally supported AI tools. It is also consistent with the NHC's broader position that rigor should be fit for purpose rather than equated with a single evidentiary method. The amount and type of evidence should be driven by the consequences of error, the degree of autonomy, the novelty of the use case, the feasibility of human review, and the degree to which performance can be reliably observed and corrected after deployment.^{9,10}

FDA would also benefit from explaining how the competency-based framework aligns with different levels of clinical confirmation. A lower-risk informational function with transparent sources, easily reversible outputs, and reliable human review may be adequately supported by rigorous benchmarking together with retrospective or shadow-deployment evidence. A patient-facing function that provides individualized action-directing guidance may warrant standardized-patient testing, prospective observation, or other evidence demonstrating that intended users understand and respond appropriately to the output. For autonomous or high-consequence functions, prospective clinical evidence may be appropriate when residual uncertainty cannot reasonably be resolved through benchmarking, simulation, or retrospective data alone.

⁹ National Health Council, "NHC Responds to NIH Artificial Intelligence Strategy RFI," July 15, 2025, <https://nationalhealthcouncil.org/letters-comments/nhc-responds-to-nih-artificial-intelligence-strategy-rfi/>.

¹⁰ Biotechnology Innovation Organization (BIO), "BIO Comments to FDA on Using Artificial Intelligence and Machine Learning in the Development of Drug and Biological Products," August 3, 2023, <https://www.bio.org/letters-testimony-comments/bio-comments-fda-using-artificial-intelligence-and-machine-learning>.

The central question should be whether the evidence package is sufficient for the intended use and clinical context rather than whether one study design is preferred in all cases.

Clinical confirmation should also consider the interaction between the device and the surrounding workflow. Performance may differ when clinicians are operating under time pressure, when staffing is limited, when a system is integrated into an electronic health record, or when a patient receives an output directly outside a clinical encounter. Automation bias, alert fatigue, confirmation bias, and overreliance on polished language can affect how otherwise accurate information is used. The NHC encourages FDA to consider these human and organizational conditions when they are material to the device's benefit-risk profile.

Benchmarks Should Reflect the Intended Population, Communication Needs, and Real-World Alternative

The proposed benchmarking elements provide a strong starting point, particularly safety-critical recognition and escalation, boundary adherence, calibration and deferral, clinical analysis, communication quality, robustness, subgroup performance, and agentic competencies. The NHC encourages FDA to treat patient comprehension and usability as substantive performance considerations rather than secondary interface issues. A clinically accurate response may still provide limited value or create risk if patients misunderstand it, cannot distinguish uncertainty from certainty, or are unable to identify the action the device is asking them to take.¹¹

Evaluation should therefore include users with varied health literacy, digital literacy, language proficiency, disability status, age, geography, and familiarity with the relevant disease area. It should also test nonstandard phrasing, colloquialisms, incomplete histories, caregiver-mediated interactions, and accessibility features where relevant. Patient organizations can help identify the terminology, scenarios, burdens, and failure modes that matter in particular disease communities. The NHC's Patient Experience Mapping Toolbox illustrates how structured patient input can identify the real-world steps, barriers, and impacts that are often invisible in conventional development processes. This input can also inform fit-for-purpose evaluation design.¹²

Patient organizations do more than recruit participants for usability testing. They can help identify the situations in which errors would be most consequential to a particular disease community, the language patients use to describe symptoms and treatment experiences, and the practical barriers that affect whether an escalation recommendation can be followed. For chronic and rare diseases in particular, patients may encounter nuanced treatment tradeoffs, fragmented specialty care, or atypical manifestations that are poorly represented in general-purpose datasets. The NHC

¹¹ Medical Device Innovation Consortium, "Patient Centered Digital Health," accessed October 6, 2026, <https://mdic.org/our-work/patient-centered-digital-health/>.

¹² Elisabeth M. Oehrlein et al., on behalf of the Patient Experience Mapping Working Group, Patient Experience Mapping Toolbox (Washington, DC: National Health Council, 2021), <https://nationalhealthcouncil.org/additional-resources/patient-experience-map/>.

encourages FDA to support early, fit-for-purpose engagement with patient and caregiver communities when benchmark scenarios, endpoints, communication criteria, and postmarket monitoring priorities are developed.^{13,14}

Comparator selection should likewise reflect the clinical question and the actual alternative available in practice. In some settings, specialist consensus may be appropriate; in others, the relevant alternative may be generalist care, delayed specialist review, no intervention, or a clinician using the device as part of a human-AI team. Requiring every GenAI system to outperform a specialist acting alone could be unnecessarily restrictive when the intended benefit is improved access in settings where specialist care is not readily available. Conversely, comparison with a median clinician may not be sufficient for a highly autonomous system making decisions with irreversible or life-threatening consequences. FDA should preserve flexibility while asking sponsors to justify why the selected comparator reflects the intended use and decision environment.

When a device is designed to augment rather than replace a clinician, FDA should also consider the performance of the human-AI team. A system that performs well in isolation may add limited value if its outputs are difficult to interpret, interrupt workflow, or encourage inappropriate reliance. Conversely, a system with imperfect standalone performance may improve care if it reliably helps clinicians recognize relevant alternatives or identify cases that might otherwise be missed. Evaluation can therefore consider whether the device changes clinician behavior, whether users appropriately accept and reject recommendations, and whether the combined system performs safely across realistic workloads and settings.

Synthetic Data Can Supplement Real-World Evidence When Fit for Purpose

The NHC recognizes the potential value of synthetic data for rare scenarios, underrepresented subgroups, stress testing, and simulation of clinically important edge cases. Synthetic inputs may be particularly useful when real patient data are sparse or prospective patient exposure would be inappropriate. At the same time, synthetic data may reproduce assumptions and limitations in the models or datasets used to generate them. FDA should therefore consider the role of synthetic data in relation to the specific evidentiary question and the availability of complementary real-world information.

The NHC encourages FDA to ask sponsors using synthetic data to describe how those data were generated, demonstrate their relevance to the intended population or deployment environment, and explain which performance questions the data can and cannot address. Synthetic evidence may be particularly useful as a supplement for stress testing and rare scenarios but should not routinely displace real patient data for evaluating subgroup performance, communication, usability, or other clinically

¹³ Medical Device Innovation Consortium, “Elevating Your RWE Strategy With Direct Patient Input: A Framework for Direct Patient Input in Real World Data and Evidence Generation,” accessed October 6, 2026, <https://mdic.org/resources/framework-direct-patient-input-rwe/>.

¹⁴ NHC, *Principles on Health AI*. See Principle 2: Patient-Centered and Informed; Principle 3: Patient Education and Empowerment; and Principle 13: Ethical Considerations.

consequential behavior. Where important subgroups are difficult to characterize premarket, a focused postmarket evidence plan may help address residual uncertainty.

The NHC also supports a fit-for-purpose statistical approach rather than a single formula for GenAI-enabled devices. Sample size, precision, and uncertainty should reflect intended use, the consequences of error, the expected frequency of safety-critical scenarios, and clinically relevant subgroup analyses. Sponsors should prespecify endpoints and analysis plans and provide sufficient evidence to characterize performance where failures could materially affect patients.

For open-ended outputs, performance assessment may need to combine several measures rather than rely on a single accuracy endpoint. Depending on intended use, relevant measures could include safety-critical failure rates, calibration, consistency across repeated runs, agreement with adjudicated clinical judgments, intended-user comprehension, time to appropriate escalation, and subgroup-specific performance. The NHC encourages FDA to promote uncertainty reporting that informs regulatory decisions. A large volume of routine interactions should not obscure rare but severe failure modes. Sponsors should also explain how ambiguous or partially correct outputs are scored and how disagreement among human adjudicators is handled.

Similarly, benchmarking and clinical-confirmation evidence need not be combined into a single estimate merely because both can be expressed numerically. Integration may be appropriate when datasets, endpoints, distributions, reference standards, and testing conditions are sufficiently comparable, and the analytical approach is prospectively justified. Where synthetic cases, retrospective records, standardized patients, and prospective encounters represent meaningfully different environments, separate reporting may provide a clearer picture of performance and uncertainty before FDA considers any combined estimate.

Independent Third Parties Can Support Evaluation While Preserving Flexibility

Qualified independent third parties may add value by maintaining sequestered benchmarks, conducting specialized testing, providing expert adjudication, or developing FDA-qualified tools. Independent infrastructure can reduce the risk of optimizing narrowly for benchmark performance and avoid unnecessary duplication across sponsors. The NHC encourages FDA to preserve flexibility in how sponsors demonstrate scientific rigor so that third-party participation does not become a de facto requirement, create avoidable development burden, or disadvantage smaller innovators. Qualifications, conflicts of interest, methods, and data-governance practices should be transparent.

The Framework Should Remain Technology-Neutral While Accounting for Architecture-Specific Risks

The NHC encourages FDA to preserve a technology- and architecture-neutral regulatory framework focused on intended use, patient risk, and demonstrated performance rather than creating separate regulatory categories for each emerging model architecture. At the same time, evidentiary expectations should reflect architecture-specific failure modes. For multimodal vision-language systems, for

example, evaluation may need to consider whether information from different modalities is integrated consistently; whether missing, corrupted, or conflicting inputs are recognized; whether errors in one modality affect downstream reasoning; and whether the system can communicate which inputs materially influenced a conclusion. Other emerging architectures may raise different questions about temporal consistency, simulation fidelity, uncertainty, or the consequences of predictions that are later used to direct or take action.

Multimodal systems illustrate why this flexibility is important. A device may perform well when text, images, laboratory values, and other inputs are evaluated separately but behave differently when those inputs conflict, arrive at different times, or are incomplete. The NHC encourages FDA to allow testing methods to reflect these foreseeable conditions and to focus on whether the complete device performs reliably in its intended environment. For generative or predictive world models, the relevant question is similarly whether model outputs translate into reliable clinical behavior rather than whether the underlying simulation appears realistic.

A technology-neutral approach can also make the framework more durable as new model types emerge. Rather than attempting to anticipate every architecture, FDA can identify the safety properties, evidence expectations, and lifecycle controls that should be demonstrated for the intended use and allow the specific testing methods to evolve with the technology.

Postmarket Monitoring Is Essential to a Total Product Lifecycle Approach

The NHC strongly supports a total product lifecycle approach for GenAI-enabled devices. Variable outputs, changing deployment environments, model updates, and dependencies on third-party foundation models make ongoing evaluation particularly important. The NHC has previously encouraged robust postmarket surveillance and continuous verification of AI-enabled devices, including audit trails, monitoring of model updates, and mechanisms to identify changes in subgroup performance and real-world reliability. This approach is also consistent with stakeholder work emphasizing ongoing quality systems, postmarket surveillance, user feedback, and adverse event monitoring as components of FDA's lifecycle framework.^{15,16}

FDA should exercise caution when deciding whether to allow greater premarket uncertainty in exchange for stronger postmarket monitoring. Such an approach may be appropriate for lower- or moderate-risk functions where errors are readily detectable and reversible, monitoring can identify deterioration promptly, and there is a reliable mechanism to intervene before harm accumulates. Greater reliance on postmarket evidence may be less appropriate for highly autonomous functions, irreversible actions,

¹⁵ National Health Council, "NHC Submits Comments on FDA Draft Guidance for AI/ML-Enabled Medical Devices," April 7, 2025, <https://nationalhealthcouncil.org/letters-comments/nhc-submits-comments-on-fda-draft-guidance-for-ai-ml-enabled-medical-devices/>.

¹⁶ Advanced Medical Technology Association (AdvaMed), "AdvaMed Responds to Congressional RFI on AI in Health Care," May 9, 2024, <https://www.advamed.org/industry-updates/news/advamed-responds-to-congressional-rfi-on-ai-in-health-care/>.

decisions with life-threatening consequences, or settings where failures are difficult to detect or attribute. The NHC encourages FDA to make these distinctions explicit so that postmarket monitoring complements, rather than substitutes for, evidence that can reasonably be generated before deployment.

Postmarket plans should be specific enough to support meaningful oversight. Relevant elements may include performance thresholds, subgroup analyses, sampling methods, reassessment frequency, triggering events, escalation procedures, and the actions expected when thresholds are crossed. Triggering events may include clinically meaningful changes to the foundation model, prompts, retrieval sources, guardrails, orchestration logic, user interface, intended population, deployment setting, or connected tools—not only formal retraining events.

A credible postmarket approach should also distinguish findings suited to routine review from signals that warrant more immediate action. Some findings may appropriately be evaluated at a regularly-scheduled review. Others—such as a material increase in a safety-critical failure, a foundation-model update that changes refusal or escalation behavior, or evidence of deteriorating performance in a clinically important subgroup—may call for prompt investigation or mitigation. The NHC encourages FDA to promote prospective planning for these response pathways to the extent practicable.

Postmarket monitoring can also support learning when important uncertainties remain at authorization, but it should not become a default substitute for feasible premarket evidence. Real-world evidence may be particularly valuable for identifying rare failure modes, changes in population mix, and performance across varied workflows. At the same time, postmarket data may be observational, incomplete, or difficult to interpret. Where FDA accepts greater premarket uncertainty, the NHC encourages clear data-quality expectations, preservation of relevant logs and version information, and a defined plan for how postmarket evidence will address the remaining uncertainty.¹⁷

Patient and Clinician Feedback Should Inform Postmarket Monitoring

Periodic re-benchmarking, clinician review, and drift monitoring are important, but they may not capture every failure mode that emerges in practice. The NHC encourages FDA to support practical mechanisms for patients, caregivers, and clinicians to report confusing, unsafe, discriminatory, or repeatedly incorrect outputs, including near misses that may not meet traditional adverse-event thresholds. Complaints, support channels, user feedback, and patient-organization input can provide useful signals and help identify patterns that structured testing may miss.

¹⁷ Medical Device Innovation Consortium, NEST Research Methods Framework: A Practical Guide to Real-World Evidence (RWE) for Medical Devices, accessed October 6, 2026, <https://mdic.org/resources/nest-research-methods-framework-a-practical-guide-to-real-world-evidence-rwe-for-medical-devices/>.

Where feasible, manufacturers should monitor whether performance changes across relevant subgroups and interaction contexts over time.¹⁸ A system can remain stable on an aggregate benchmark while becoming less reliable for a smaller population, when responding to a particular language pattern, or in a care setting experiencing distribution shift. These concerns are consistent with broader clinical and stakeholder work identifying dataset limitations, embedded disparities, accountability, reliability, cybersecurity, and the ethics of patient data as important considerations for AI in health care.¹⁹

Subgroup monitoring should be clinically motivated and sufficiently granular to identify meaningful disparities while recognizing the limitations of small datasets. Relevant subgroups will vary by intended use and may be defined by age, disability, language, sex, race and ethnicity, geography, disease severity, comorbidities, socioeconomic barriers, or other factors that affect available inputs or interpretation of outputs. FDA should encourage sponsors to explain why selected subgroup analyses are relevant and how sparse data will be addressed. When premarket evidence cannot reliably characterize a small population, that limitation should be clearly disclosed and, where appropriate, paired with a focused postmarket plan.

Equity considerations should include access and usability as well as algorithmic performance. A device may perform similarly across demographic groups but still have unequal effects if it assumes reliable broadband, high digital literacy, English proficiency, ready access to follow-up care, or the ability to pay for recommended services. For patient-facing GenAI devices, FDA should consider whether foreseeable access barriers affect the consequences of an output and whether communication and escalation pathways are workable for the populations the device is intended to serve. These considerations are particularly important when a device is expected to expand access in rural, underserved, or specialist-shortage settings.

Machine-Based Supervisory Agents May Support Monitoring but Require Independent Evaluation

Machine-based supervisory agents may help scale monitoring by detecting anomalous outputs, policy violations, drift, or changes in conversational behavior. Their use, however, does not eliminate the need to evaluate the monitoring function itself. Supervisory systems may share data sources, architecture, assumptions, or failure modes with the device they monitor. The NHC encourages FDA to clarify expectations for validating supervisory agents, characterizing false-negative and false-positive rates, defining when human review is triggered, and periodically assessing whether the monitoring approach continues to identify known and emerging failure modes.

¹⁸ NHC, *Principles on Health AI*. See Principle 4: Equity and Accessibility; Principle 7: Bias Mitigation; Principle 10: Continuous Improvement; and Principle 11: Safety and Effectiveness.

¹⁹ American Heart Association, “Pragmatic Approaches to the Evaluation and Monitoring of Artificial Intelligence in Health Care,” November 10, 2025, <https://professional.heart.org/en/science-news/pragmatic-approaches-to-the-evaluation-and-monitoring-of-artificial-intelligence-in-healthcare>.

Shared Postmarket Roles Should Preserve Clear Manufacturer Accountability

The NHC agrees that clinicians, health systems, professional societies, standards organizations, patient organizations, payers, and other stakeholders can contribute to postmarket monitoring and responsible deployment. Health systems can identify local workflow problems and distribution shifts; professional societies can help define clinically appropriate performance expectations; patient organizations can identify potential patient harm and communication concerns; and standards organizations can support common testing methods. These complementary roles can strengthen oversight while preserving the device manufacturer's responsibility for the safety and effectiveness of the marketed device.

Clear expectations are especially important when devices depend on third-party foundation models. The NHC encourages FDA to consider how contractual and technical mechanisms can support timely notice of material model changes, version identification, access to safety-relevant documentation, and the ability to evaluate or temporarily suspend affected functionality when necessary. Audit logs can also support reconstruction of the model version, prompt configuration, retrieval source, and tool environment associated with a consequential output. If a third-party dependency cannot be adequately monitored or managed, that limitation should be considered in the device's benefit-risk assessment and change-control approach.

The level of visibility and control expected of the device manufacturer should be proportionate to the role the foundation model plays in the regulated function. At a minimum, manufacturers should be able to identify the model version in use, detect material changes, understand known limitations relevant to intended use, and evaluate whether an update changes safety-critical behavior. For higher-risk devices, advance notice, change documentation, access to evaluation information, or continued access to a previously validated model version may be important safeguards. FDA should consider these dependencies when evaluating the adequacy of a sponsor's change-control strategy.

Patients and clinicians should also understand who is responsible for addressing problems. A complex chain of model developers, device manufacturers, health systems, integrators, and clinicians can make it difficult for users to know where to report a problem or who will investigate it. The NHC encourages FDA to promote clear labeling and reporting pathways and to maintain a clear point of accountability for coordinating investigation and corrective action, including when a third-party component contributes to the problem.

Predetermined Change Control Plans Can Support Responsible Adaptation

The NHC supports the use of predetermined change control plans (PCCPs) where future modifications can be described with enough specificity to define the methods, acceptance criteria, and boundaries for making changes without undermining safety or effectiveness. FDA's existing PCCP guidance provides an important foundation for adaptive AI-enabled device software functions. For GenAI systems, the challenge will

be to define change categories by function so that they remain useful even when the exact future implementation cannot be predicted.^{20,21}

A GenAI PCCP could, for example, prespecify classes of foundation-model updates, retrieval-source changes, prompt or guardrail modifications, or performance-tuning changes, together with the re-benchmarking and clinical evidence appropriate for each class. The NHC encourages FDA to provide sufficient guardrails around broadly defined change categories so that flexibility does not reduce clarity regarding when additional testing or Agency review is warranted. As future modifications become more difficult to characterize in advance, clear post-change testing, version control, monitoring, and notification thresholds become increasingly important.

The extent of reassessment should reflect the expected clinical impact of a modification. Changes limited to formatting, latency, or nonclinical user-interface features may be appropriately managed within a quality system when they do not alter the meaning, directiveness, or reliability of outputs. Changes affecting the underlying model, clinical knowledge sources, retrieval logic, safety guardrails, escalation thresholds, autonomy, or intended populations may warrant more substantial re-evaluation. FDA should encourage sponsors to define objective criteria for determining when a change can be managed within the quality system, when PCCP testing is appropriate, and when a new premarket submission may be needed.

Foundation Model Device Master Files Could Improve Efficiency if They Are Current, Standardized, and Useful

The NHC supports FDA's exploration of voluntary Foundation Model Device Master Files (MAFs). A well-designed MAF program could reduce duplicative submissions, give review teams a more consistent understanding of widely used foundation models, and provide a mechanism for model developers to share confidential safety-relevant information that individual device sponsors may not otherwise possess. This could be particularly valuable as more medical devices rely on a smaller number of general-purpose foundation models.

To be useful, Foundation Model MAFs would need to be versioned and kept current. Relevant content could include training-data provenance at an appropriate level of detail, architecture and supported modalities, known limitations and failure modes, evaluation results relevant to health care, subgroup performance, refusal and escalation behavior, safety guardrails, update cadence, material-change notification practices, cybersecurity considerations, audit-log availability, and the circumstances under which

²⁰ U.S. Food and Drug Administration, Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions: Guidance for Industry and Food and Drug Administration Staff (August 2025), <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence>.

²¹ Medical Device Innovation Consortium, Predetermined Change Control Plans (PCCPs) for Artificial Intelligence (AI) and Machine Learning (ML)-Enabled Medical Devices, September 1, 2024, <https://mdic.org/wp-content/uploads/2024/09/MDIC-AI-ML-Framework-v2.pdf>.

prior evaluation may no longer apply. The program should also encourage standardized model or system cards that provide a common structure for this information. The NHC has previously pointed to model-card frameworks as an example of infrastructure that can support transparent, reproducible evaluation.²²

Voluntary MAF participation should not become a de facto prerequisite that disadvantages device developers whose foundation-model providers do not participate. At the same time, the absence of a MAF should not reduce the evidence needed to support the safety and effectiveness of the final device. The NHC encourages FDA to preserve sponsor responsibility for understanding and managing safety-relevant dependencies on the underlying model.

For the MAF program to remain useful over time, FDA should consider expectations for maintenance and notification. A file that accurately describes a foundation model at initial submission but is not updated after significant changes may provide limited value to downstream device sponsors and reviewers. Participating model developers could identify the versions covered by the file, define material updates, and provide timely revisions or notices when safety-relevant characteristics change. Standardized fields for update cadence, deprecation policies, known health care limitations, and availability of prior versions may also improve the usefulness of MAFs.

Because participation would be voluntary, FDA should also consider incentives that make MAFs useful to model developers without creating the impression that the underlying foundation model has been authorized for medical use. Predictable gains in review efficiency, clearer expectations for information commonly requested across device submissions, and confidential handling of proprietary technical information may encourage participation. FDA should also make clear that each sponsor remains responsible for validating the complete device configuration, including prompts, retrieval systems, guardrails, interfaces, tools, and the deployment environment.

Agentic GenAI Warrants Additional Risk-Proportionate Safeguards

Agentic GenAI warrants heightened attention because risk can arise from a sequence of individually plausible actions rather than a single output. An agent may retrieve data, interpret results, call tools, generate orders, communicate with patients, or control another device across multiple steps. Small errors can compound, and opportunities for human review may diminish as the system becomes more autonomous. The NHC supports FDA's inclusion of agentic competencies within the proposed benchmarking framework and encourages additional evaluation of cumulative failure, tool misuse, permission boundaries, and recovery from intermediate errors.

As an agentic system gains authority to take consequential actions, acceptance criteria and safeguards should become more stringent. Depending on intended use and risk, safeguards may include clearly defined permissions and scope, least-privilege access to data and tools, and meaningful human authorization before irreversible or high-

²² National Health Council, "NHC Responds to NIH Artificial Intelligence Strategy RFI," July 15, 2025, <https://nationalhealthcouncil.org/letters-comments/nhc-responds-to-nih-artificial-intelligence-strategy-rfi/>.

consequence actions. Other safeguards may include reliable detection of erroneous tool outputs, auditable action logs, fail-safe states, rollback or interruption mechanisms, and resistance to prompt injection or malicious content introduced through retrieved information or connected tools. Evaluation should also consider whether the system recognizes when a planned sequence has moved outside its intended use even if individual steps appear permissible.

These considerations reinforce the importance of meaningful human oversight. When human review serves as a safeguard, the reviewer should have sufficient time, information, authority, and understanding to identify and interrupt an unsafe action.

Agentic systems also raise questions about cumulative risk. A system may make several individually plausible choices—selecting a data source, summarizing information, choosing a tool, setting a parameter, and initiating an action—that collectively lead to an unsafe result. The NHC encourages FDA to evaluate full task sequences, error propagation, recovery after an incorrect intermediate step, and the ability to stop when uncertainty exceeds an appropriate threshold. Acceptance criteria should consider not only the final action but also whether the sequence was auditable, appropriately constrained, and recoverable.

Human oversight should also be evaluated in the context of the actual workflow. A requirement for clinician approval may provide limited protection if a system presents numerous recommendations at high speed, supporting information is unavailable, or the workflow makes rejection difficult. The NHC encourages FDA to consider whether oversight is timely, informed, and operationally realistic. For higher-consequence actions, affirmative confirmation at a meaningful checkpoint may be more appropriate than relying on the theoretical availability of a reviewer after the fact.

Transparency, Trust, and Patient Engagement Should Remain Cross-Cutting Priorities

Although the discussion paper appropriately focuses on regulatory science, successful implementation will also depend on whether patients and clinicians understand and trust the devices they use. The NHC encourages FDA to preserve transparency as a cross-cutting expectation. Patients should have access to understandable information about when a regulated GenAI function is materially shaping information or decisions, and what the tool is intended to do. This information should also explain the tool's limitations, when human review is available or recommended, and how significant changes may affect use.²³ Technical disclosure alone has limited value if the information is not usable by the people making decisions.

²³ NHC, *Principles on Health AI*. See Principle 3: Patient Education and Empowerment; Principle 5: Trustworthiness; Principle 6: Comprehensive Transparency; Principle 9: Human Collaboration; and Principle 12: Accountability.

The NHC has repeatedly emphasized early and sustained patient engagement in AI development and oversight.²⁴ Patients and caregivers can help identify which errors are most consequential, what levels of uncertainty may be acceptable, which explanations are understandable, and what escalation pathways are practical. Meaningful engagement can inform not only interface testing but also intended use, benchmark scenarios, communication criteria, postmarket signals, and change-management priorities.

FDA can make this expectation more concrete by encouraging sponsors to describe patient engagement in regulatory submissions when it materially informed device design or evaluation. Relevant information may include which patient and caregiver populations were engaged, when engagement occurred, the questions considered, how participants were selected, and what changed as a result. This need not become a rigid procedural requirement. Rather, documenting how patient input influenced the development program can help FDA assess whether engagement was timely, fit for purpose, and relevant to the device's foreseeable benefits and risks.

Transparency should likewise be tailored to the decision being made and the needs of the intended audience. A patient may not need detailed information about model architecture to use a lower-risk tool safely. The patient will still need to understand that an individualized recommendation is generated by AI, that the system can be wrong, what information it used, whether a clinician reviewed the output, and how to seek review or correction. Clinicians may need additional information about validation populations, known failure modes, changes across versions, and circumstances in which the system should not be relied upon. The NHC encourages FDA to support layered communication that makes the most relevant information usable for different audiences.

These expectations are also consistent with perspectives across the patient and medical technology communities emphasizing representative evidence, transparency, meaningful human oversight, postmarket surveillance, privacy, and safe use. These perspectives reinforce a central point: innovation and patient protection are not opposing objectives. A predictable, evidence-based, patient-centered framework can increase confidence in beneficial technologies and help distinguish them from tools that have not demonstrated reliable performance.^{25,26}

²⁴ NHC, *Principles on Health AI*. See Principle 1: Patient Health Benefit; Principle 2: Patient-Centered and Informed; Principle 3: Patient Education and Empowerment; and Principle 13: Ethical Considerations.

²⁵ Alzheimer's Association, "Technology Safety for Older Adults," accessed October 6, 2026, <https://www.alz.org/help-support/caregiving/safety/technology-safety-older-adults>.

²⁶ Advanced Medical Technology Association (AdvaMed), The Role of Artificial Intelligence (AI) in Healthcare, November 11, 2024, <https://www.advamed.org/member-center/resource-library/the-role-of-artificial-intelligence-ai-in-healthcare/>.

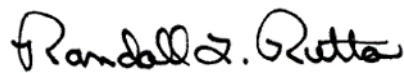
Conclusion

The NHC appreciates FDA's decision to seek early public input on the regulation of GenAI-enabled medical devices. The discussion paper provides a useful foundation by recognizing that generative systems may require different approaches to risk assessment, evidence generation, change control, and postmarket monitoring from those used for traditional software. It also preserves risk-based and least-burdensome principles. The NHC supports continued development of the proposed risk and competency-based frameworks, with particular attention to patient context, real-world comprehension, subgroup performance, clinically meaningful confirmation, transparency, and clear accountability across the total product lifecycle.

As FDA develops future policy, the NHC encourages the Agency to continue engaging patients and patient organizations as active partners in defining what safe, effective, and trustworthy GenAI looks like in practice. A durable framework should support innovation while maintaining evidence expectations that are proportionate to risk and responsive to changes in models, populations, and deployment environments. A patient-centered, fit-for-purpose approach can help beneficial GenAI-enabled devices reach patients while building the confidence needed for responsible adoption.

Please do not hesitate to contact Jennifer Dexter, Senior Vice President, Policy & External Affairs, or Shion Chang, Assistant Vice President, Policy & Regulatory Affairs, if you or your staff would like to discuss these comments in greater detail.

Sincerely,

A handwritten signature in black ink that reads "Randall L. Rutta". The signature is written in a cursive, flowing style.

Randall L. Rutta
Chief Executive Officer