

AMERICAN ASSOCIATION OF
NEUROLOGICAL SURGEONS
KATIE O. ORRICO, CEO
5550 Meadowbrook Drive
Rolling Meadows, IL 60008
Phone: 888-566-AANS
Fax: 847-378-0600
info@aans.org



CONGRESS OF
NEUROLOGICAL SURGEONS
REGINA SHUPAK, CEO
10 North Martingale Road, Suite 190
Schaumburg, IL 60173
Phone: 877-517-1CNS
FAX: 847-240-0804
info@cns.org

President
E. ANTONIO CHIOCCA, MD, PHD
Boston, Massachusetts

President
MARTINA STIPLER, MD
Boston, Massachusetts

VIA ELECTRONIC TRANSMISSION

October 19, 2026

Kyle Diamantas, JD
Acting Commissioner
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

Dear Acting Commissioner Diamantas,

On behalf of the American Association of Neurological Surgeons (AANS) and the Congress of Neurological Surgeons (CNS), representing more than 4,000 neurosurgeons nationwide, we appreciate the opportunity to submit comments on the Food and Drug Administration's (FDA) *Considerations for the Regulation of Generative AI-Enabled (GenAI) Medical Devices* discussion paper, published on August 18, 2026.

The discussion paper addresses several issues of significant importance to neurosurgical patients and providers. In particular, the AANS and CNS comment on the assessment of risk, competency-based premarket evaluation, and postmarket monitoring of GenAI devices.

In summary, the AANS and CNS recommend that the FDA:

- Refine its proposed risk framework by evaluating the accuracy, reliability, and potential clinical consequences of GenAI-generated outputs, rather than relying solely on the type of activity the GenAI system performs;
- Move beyond the "non-directive" versus "action-directing" framework and instead assess GenAI systems based on the degree to which their outputs influence clinical decisions, with appropriate safeguards;
- Consider requiring certain safeguards, especially for patient-facing GenAI functions, that are activated when a model's confidence is low or when the system detects a heightened risk of hallucination;
- Encourage developers to incorporate risk-based safeguards that account for the complexity of the clinical question and the clinical expertise of the user to reduce the risk of misinterpretation;

- Assess GenAI risk based on the overall impact of multi-turn conversations and the foreseeable clinical actions they influence, not individual responses viewed in isolation;
- Evaluate GenAI-enabled devices in the context of the clinical decisions they influence and the downstream consequences of referral-related errors;
- Assess transparency as part of Clinical Proficiency, with expectations proportionate to a device's intended use and risk profile, to ensure users can understand, evaluate, and appropriately rely on GenAI-generated outputs;
- Tailor premarket evidence expectations to the intended use, risk profile, level of autonomy, and potential impact of the technology on clinical decision-making and patient outcomes, ensuring that regulatory requirements are appropriately aligned with risk; and
- Establish clear, risk-based Predetermined Change Control Plan (PCCP) policies that allow appropriately validated lower-risk GenAI updates to proceed without new premarket review while maintaining oversight of higher-risk modifications.

The AANS and CNS appreciate the FDA's consideration of these comments and look forward to continued engagement on policies that accelerate patient access to innovative and safe medical devices that improve long-term health outcomes and strengthen quality of care. As the FDA continues to develop its regulatory approach to GenAI-enabled and other emerging medical devices, policies must support innovation while ensuring appropriate standards for safety, effectiveness, and transparency. The AANS and CNS respectfully offer the following responses to the specific questions posed in the discussion paper.

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 AANS and CNS appreciate the proposed two-axis framework as a starting point for evaluating GenAI-related risk in medical devices. However, the FDA should avoid treating the framework as a rigid or comprehensive measure of risk. While the nature of a device's activity is an important consideration, risk is also significantly influenced by the quality, reliability, and clinical appropriateness of the GenAI-generated output. As a result, devices that fall within the same category of activity may present substantially different risk profiles depending on how accurately and consistently the underlying GenAI system performs.

In addition, the potential consequences of an incorrect output may be significant regardless of whether the GenAI system is providing information, conducting analysis, or generating a specific recommendation. An inaccurate informational output, for example, may shape clinical decision-making in ways that ultimately affect patient care just as much as the wrong recommendation. Accordingly, risk assessments should consider not only the type of output generated, but also the likelihood and potential impact of errors in the clinical context in which the technology is used.

These considerations are particularly important for GenAI systems, which may perform less reliably in rare, complex, or atypical patient scenarios.¹ In such cases, the system may generate outputs or recommendations that appear plausible and clinically sound, but are inappropriate for the specific

¹ Thomas Maddox et al., *Risks of Generative Artificial Intelligence in Health and Medicine*, in *Generative Artificial Intelligence in Health and Medicine: Opportunities and Responsibilities for Transformative Innovation* (National Academy of Medicine 2025)

circumstances. Because these types of errors may be difficult for users to detect, the FDA should recognize the limitations of GenAI in edge cases and account for the potential for clinically consequential but seemingly credible outputs when evaluating risk.

For example, a GenAI-enabled neurosurgical decision support tool may be used to summarize imaging findings and relevant clinical information for a patient with a brain tumor, such as a glioma. Although the tool's function may be characterized as primarily "informational" under the framework, an inaccurate summary that omits tumor involvement near eloquent brain regions or mischaracterizes the extent of disease could influence surgical planning and patient counseling.² In this circumstance, the clinical risk stems not merely from the type of activity performed by the GenAI system, but from the accuracy and reliability of the output itself.

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?

The discussion paper appropriately acknowledges that "non-directive" and "action-directing" categories may not be binary but instead exist along a continuum. This recognition is important because GenAI-generated information is often used in ways that do not fit neatly into either category. In practice, many GenAI outputs combine elements of information sharing, risk assessment, and clinical decision support. For example, an estimated risk score or patient-specific summary may not explicitly recommend a course of action, yet such outputs can meaningfully influence clinical judgment, care planning, and patient management decisions.

The FDA has recognized that software may "support clinical decision-making" even when it does not direct a specific action, and researchers have observed that risk-assessment and decision-support tools are used to augment clinician decision-making and can affect care delivery and patient outcomes.³ Conversely, an output that includes a specific recommendation may be only one factor among many a healthcare professional considers. As a result, the practical impact of GenAI output cannot be determined solely by whether it contains an explicit instruction or recommendation.

Consistent with CDRH's recognition that GenAI outputs may fall along a continuum of influence, the AANS and CNS recommend that FDA evaluate risk based on the degree to which an output is likely to shape clinical decision-making and patient management, rather than relying primarily on whether the output explicitly recommends an action. A risk-based framework that accounts for varying levels of clinical significance, user reliance, and decision-support functionality would better reflect real-world clinical workflows and provide a more accurate basis for assessing the potential risks and benefits of GenAI-enabled technologies. This approach would also help ensure that regulatory oversight remains proportionate to the actual impact of GenAI outputs on healthcare decision-making.

Healthcare providers may rely on both generalized and more directive outputs when making decisions.

² Sarah K. Koning et al., *Preoperative Assessment of Tumor Eloquence and Resectability: An International Survey*, *Journal of Neuro-Oncology* (2025), <https://doi.org/10.1007/s11060-025-05067-0>

³ U.S. Food & Drug Admin., *Clinical Decision Support Software Frequently Asked Questions (FAQs)*, <https://www.fda.gov/medical-devices/software-medical-device-samd/clinical-decision-support-software-frequently-asked-questions-faqs>

For example, clinicians may act on a predicted risk of disease progression or a probability of adverse events in the same way they might act on a recommendation to order additional testing or modify a treatment regimen. The key consideration is not whether the output is framed as information or advice, but rather the degree to which it informs and influences clinical decision-making in a particular context. A binary regulatory distinction between "non-directive" and "action-directing" outputs may therefore fail to capture the range of ways GenAI systems affect clinical care.

As GenAI systems increasingly generate outputs informed by patient-specific information and present it in more actionable ways, transparency and context become critical safeguards. GenAI-generated outputs should clearly identify the source and date of material information and distinguish documented facts from model-generated conclusions.⁴ In addition, encounter-specific context, including the clinical setting, visit type, and intended user role, should inform whether support is presented and how it is communicated. These measures would enhance trustworthiness and help ensure that clinicians can appropriately evaluate GenAI-generated information while maintaining responsibility for clinical decision-making.

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?

A major concern with using GenAI in patient-facing healthcare applications is the risk of model "hallucination," where a GenAI system generates information, recommendations, or conclusions that appear authentic and reliable but are inaccurate or entirely fabricated.⁵ Unlike healthcare professionals, GenAI systems may provide highly confident responses even when clinical uncertainty exists, or no reliable answer can be supported by the underlying evidence. While patients increasingly access and interpret complex health information, they may not always possess the clinical training necessary to independently evaluate the accuracy or reliability of GenAI-generated outputs. As a result, inaccurate information may influence decisions regarding whether and when to seek medical attention.

At the same time, the FDA should avoid adopting an approach that assumes patients are incapable of meaningfully engaging with health information. Patient-facing GenAI technologies have the potential to improve health literacy, increase engagement in care, facilitate shared decision-making, and expand access to information for individuals who may otherwise face barriers to obtaining timely clinical guidance.^{6,7} The appropriate regulatory objective is to enable patient access to beneficial information while ensuring that device design appropriately accounts for situations in which patients may lack the expertise necessary to independently verify an output.

⁴ Siru Liu, PhD, Josh F. Peterson, MD, MPH, and Adam Wright, PhD, "How Generative AI Should Transform Clinical Decision Support," *JAMA* 336, no. 7 (2026), <https://jamanetwork.com/journals/jama/fullarticle/2851998>

⁵ Rami Hatem, Brianna Simmons, and Joseph E. Thornton, "A Call to Address AI 'Hallucinations' and How Healthcare Professionals Can Mitigate Their Risks," *Cureus* 15, no. 9 (September 5, 2023): e44720, <https://doi.org/10.7759/cureus.44720>

⁶ Tracie Risling et al., Defining Empowerment and Supporting Engagement Using Patient Views From the Citizen Health Information Portal: Qualitative Study, 6 *JMIR Med. Informatics* e43 (2018), <https://pmc.ncbi.nlm.nih.gov/articles/PMC6231721/>

⁷ Samira Abbasgholizadeh Rahimi et al., *Application of Artificial Intelligence in Shared Decision Making: Scoping Review*, 10 *JMIR Med. Informatics* e36199 (2022), <https://pmc.ncbi.nlm.nih.gov/articles/PMC9399841/>

To that end, the FDA should consider a risk-based framework for patient-facing GenAI functions that places particular emphasis on the nature of the information provided, the potential consequences of acting on that information, and the ability of users to recognize when additional clinical input is warranted. Higher-risk outputs are likely to include those that recommend treatment options, suggest changes to medications, interpret clinical test results, or otherwise influence decisions that could delay or alter medical care. Lower-risk outputs may include educational content, administrative assistance, or provision of general health information that does not purport to direct individual clinical decision-making.

To mitigate risks associated with patient-facing GenAI systems, the AANS and CNS recommend the FDA consider requiring certain safeguards that are activated when a model's confidence is low or when the system detects a heightened risk of hallucination. These mechanisms could include prompting patients to seek advice from a qualified healthcare professional, directing users to validated sources of information, or escalating questions to human review when appropriate. These safeguards would help ensure that GenAI tools serve as supplements to, rather than replacements for, clinical judgment and human expertise in situations where patient safety may be at risk.

The risks associated with GenAI hallucinations are particularly concerning in specialties such as neurosurgery, where inaccurate recommendations can have profound consequences for patient safety. For example, a patient using a GenAI tool to understand symptoms such as progressive weakness, severe headaches, or changes in vision may receive an overly confident recommendation that minimizes the need for urgent evaluation, despite these symptoms potentially reflecting serious neurological conditions requiring immediate medical attention. Research evaluating large language models in neurosurgical settings has demonstrated that, while newer models can perform well on specialty-specific examinations, they can still generate incorrect information and hallucinated responses, especially when faced with complex clinical reasoning tasks.⁸

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

The distinction between generalist and specialist physicians should play a meaningful role in FDA's assessment of risk for HCP-facing GenAI-enabled functions. Risk is not determined solely by the accuracy of a GenAI-generated output, but also by the clinical expertise required to appropriately interpret and act upon that output. A recommendation or finding that can be readily evaluated by a specialist may present substantially greater risk when presented to a clinician who lacks specialized training in the relevant area of practice.

GenAI-generated explanations and recommendations do not benefit all users equally. Research suggests that explanations and recommendations generated by GenAI systems can benefit experts while misleading less experienced users, underscoring the importance of tailoring outputs to the expertise of the intended audience.⁹ The FDA should evaluate GenAI-enabled devices not only based on their intended use and technical performance, but also based on the characteristics, training, and expected competencies of the intended users.

⁸ Ahmed Alim et al., "A Descriptive Study Based on the Comparison of ChatGPT and Evidence-Based Neurosurgeons," *Cureus* 15, no. 8 (2023): e43066, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10495632/>.

⁹ Massachusetts Institute of Technology News, "The Benefits of Medical AI Assistance Vary Based on User Expertise," *MIT News*, August 4, 2026, <https://news.mit.edu/2026/medical-ai-assistance-benefits-vary-based-on-user-expertise-0804>

To mitigate these risks, the FDA should encourage developers to incorporate risk-based safeguards that account for the complexity of the clinical question and the expertise of the user. When such thresholds are exceeded, devices should provide their output alongside a recommendation for specialist consultation or additional evaluation to help ensure appropriate clinical oversight and reduce the risk of misinterpretation.

The manner in which information is presented should also vary depending on the intended user.¹⁰ Generalist-facing outputs may require additional context, plain-language explanations, and clear recommendations regarding next steps. In contrast, specialist-facing outputs may appropriately include more detailed clinical information, supporting evidence, and technical data regarding how conclusions were reached. User interface design should support safe interpretation by presenting information at a level that matches the training and responsibilities of the user. In addition, transparency regarding the underlying data sources, potential sources of bias, and the methods used to train and validate the system is particularly important for expert users who must evaluate the reliability and applicability of GenAI-generated recommendation.

Consideration should also be given to establishing access controls that reflect varying levels of clinical expertise and accountability. GenAI-enabled devices could incorporate distinct log-on requirements for specialists and general physicians, with corresponding permissions and responsibilities. This approach would create traceable records of user interactions and help ensure that the most consequential clinical decisions receive appropriate professional oversight.

As GenAI systems become increasingly integrated into clinical decision-making, questions remain regarding liability when a clinician follows a GenAI recommendation and harm occurs, as well as when a clinician exercises independent judgment and reaches a different conclusion than the GenAI system. The FDA should acknowledge these unresolved issues and work collaboratively with stakeholders to ensure that regulatory approaches promote physician oversight without creating uncertainty regarding accountability among clinicians, developers, healthcare institutions, and other parties involved in the deployment of GenAI-enabled technologies.

Finally, post-market monitoring is essential for identifying risks that may not be apparent during premarket evaluation.¹¹ Developers and regulators should have mechanisms to evaluate, in real time when possible, how GenAI models respond to different types of questions and clinical scenarios once deployed in practice. Attention should be paid to whether performance varies between specialist and non-specialist users, whether certain outputs are subject to misinterpretation, and whether patterns of overreliance emerge in practice. This monitoring will be especially important for adaptive or continuously learning systems, where real-world performance may evolve over time and require corrective action to protect patient safety.

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?

In the case of multi-turn conversational systems, risk assessment should extend beyond evaluation of prompts and responses in isolation. The FDA should consider the entirety of a realistic conversational

¹⁰ U.S. Food and Drug Administration (FDA), *Human Factors Considerations*, available at <https://www.fda.gov/medical-devices/human-factors-and-medical-devices/human-factors-considerations>

¹¹ Challenges of Implementing LLMs in Clinical Practice: Perspectives, PubMed Central (PMC), <https://pmc.ncbi.nlm.nih.gov/articles/PMC12429116/>

trajectory and how successive exchanges may incrementally shape a patient's understanding of their condition, perception of risk, and subsequent healthcare decisions. In many cases, the overall effect of a conversation may be more consequential than any single response viewed independently.

This consideration is particularly important because the distinction between "non-directive" and "action-directing" information may not remain clear over the course of a conversation. Individual responses that appear educational or informational when viewed in isolation may, when combined across multiple exchanges, steer a user toward a specific clinical conclusion or course of action. As a result, risk should be assessed based on the overall impact of the interaction and the foreseeable actions that could result from the cumulative conversation, rather than the content of any single output.

This approach is consistent with FDA's Total Product Lifecycle framework for AI-enabled devices, which recognizes that safety and effectiveness must be evaluated and managed throughout the lifecycle of an AI system, including how performance may change in real-world use.¹² Similarly, the National Institute of Standards and Technology (NIST) AI Risk Management Framework emphasizes that AI risks must be evaluated within their specific context of use and notes that human-AI interactions can vary significantly depending on how users engage with the system.¹³ NIST specifically highlights that human-AI interactions can amplify biases, influence downstream decision-making, and produce outcomes that differ from the impact of either the human or AI alone.

Regulatory evaluations should incorporate testing methodologies that examine multi-turn conversations, including scenarios in which the system transitions from providing general health information to supporting treatment-related decisions. Evaluating performance across these conversations would provide a more accurate assessment of safety and effectiveness than assessing isolated prompts and responses alone. Regulatory evaluation should include representative multi-turn scenarios that assess whether cumulative conversational exchanges effectively guide users toward clinical decisions, even when no single response is independently directive.

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?

Accurate referral, ordering additional testing, or urgent specialist consultation recommendations are critical to patient safety, particularly for GenAI-enabled devices intended to support clinical triage, diagnosis, or care pathway decisions. When evaluating these technologies, the FDA should recognize that both over-escalation and under-escalation can result in meaningful patient harm.

Excessive or inappropriate referrals may contribute to unnecessary emergency department visits, specialist consultations, diagnostic procedures, and follow-up testing.¹⁴ These downstream effects can increase healthcare costs, burden patients and caregivers, and further strain already limited healthcare resources. In some settings, high false-positive rates may also contribute to clinician alert fatigue and reduce confidence in GenAI-enabled tools, potentially limiting their effectiveness in clinical practice.

¹² U.S. Food & Drug Admin. (FDA), *Good Machine Learning Practice for Medical Device Development: Guiding Principles* (2021), <https://www.fda.gov/medical-devices/artificial-intelligence-enabled-medical-devices/good-machine-learning-practice-medical-device-development-guiding-principles>.

¹³ National Institute of Standards and Technology (NIST), *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1 (Jan. 2023), <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>.

¹⁴ Bertalan Meskó & Eric J. Topol, *Guiding Principles for the Responsible Development of Artificial Intelligence Tools for Healthcare*, 6 NPJ Digit. Med. 47 (2023), <https://pmc.ncbi.nlm.nih.gov/articles/PMC10066953/>

These risks may be particularly consequential in time-sensitive neurological conditions, where even modest delays in treatment can have lasting effects on patient outcomes. For example, a GenAI-enabled imaging tool designed to identify suspected intracranial hemorrhage on head CT and prioritize those cases for urgent review may influence escalation-of-care decisions.¹⁵ If the device exhibits a high false-positive rate, it could lead to over-escalation of care, including unnecessary neurosurgical consultations, patient transfers to care centers, additional diagnostic testing, or emergency department resource utilization for patients who do not ultimately require intervention. However, if the device fails to identify a true hemorrhage, under-escalation of care could occur, resulting in delayed neurosurgical evaluation, delayed treatment, neurological deterioration, or missed opportunities for timely intervention. In either scenario, the clinical impact extends beyond device accuracy metrics and may directly affect patient outcomes, clinician workflows, and healthcare resource utilization.

Accordingly, the FDA should evaluate GenAI-enabled devices in the context of the clinical decisions they influence and the downstream consequences of referral-related errors. Consistent with FDA's risk-based approach and the International Medical Device Regulators Forum Software as a Medical Device clinical evaluation framework, assessment of safety and effectiveness should consider whether a device's ability to correctly identify patients with and without the condition is appropriately balanced for the intended use, clinical setting, patient population, and potential clinical impact of false-positive and false-negative results.¹⁶

In addition, manufacturers should be encouraged to evaluate and monitor referral patterns throughout the product lifecycle. Pre-market studies and post-market surveillance activities should assess whether the technology is producing unintended effects on patient outcomes, clinician workflows, healthcare utilization, or access to care. Ongoing monitoring of referral behavior can help identify performance drift, emerging biases, or other real-world impacts that may not be initially apparent, which supports the safe and effective use of GenAI-enabled medical devices over time.

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?

The AANS and CNS recommend that the FDA include a transparency-focused consideration within the Clinical Proficiency category. In addition to evaluating the clinical appropriateness and reliability of GenAI-generated outputs, the FDA should assess whether a device provides users with meaningful information regarding the basis for those outputs. As appropriate to the device's intended use, risk profile, and technical design, this could include identifying the sources, references, evidence, data categories, or other inputs that informed a response or recommendation.

Transparency about how GenAI-generated content is created and what sources it is based on is important for safe and effective clinical care.¹⁷ Clinicians must be able to evaluate not only what a device recommends, but also why it reached that recommendation and whether the rationale is appropriate for a particular patient. Providing insight into the sources or rationale can help users assess the reliability

¹⁵ Grobe AF, Mayer SA, Kallmes DF, et al. *Real-world performance evaluation of a commercial deep learning model for intracranial hemorrhage detection*. Front Neurol. 2023;14:1098377. Available at: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12823562/>

¹⁶ U.S. Food and Drug Administration (FDA), Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Discussion Paper and Request for Feedback (Apr. 2, 2019), <https://www.fda.gov/media/122535/download>

¹⁷ U.S. Food & Drug Admin., Health Canada & Medicines & Healthcare Products Regul. Agency, Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles (2024), <https://www.fda.gov/medical-devices/artificial-intelligence-enabled-medical-devices/transparency-machine-learning-enabled-medical-devices-guiding-principles>

and relevance of outputs, identify potential gaps, limitations, or biases, and independently verify information when necessary.

Greater transparency can also strengthen clinician oversight and support appropriate human judgment in patient care. This consideration may be particularly important for GenAI-enabled tools used in neurosurgery. For example, AI technologies have already demonstrated the ability to detect hydrocephalus requiring treatment from neuroimaging with performance comparable to expert neuroradiologists.¹⁸ In these circumstances, clinicians would benefit from understanding whether the device's output was driven primarily by imaging features, published clinical evidence, patient-specific factors, clinical guidelines, or other data sources. Such transparency would enable neurosurgeons to independently evaluate the reliability and appropriateness of the recommendation in the context of an individual patient's care.

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 AANS and CNS do not support a blanket position that reduced premarket evidence is always appropriate for GenAI-enabled medical devices. The degree of premarket evidence should remain proportionate to the intended use, risk profile, level of autonomy, and potential impact of the technology on patient care. For certain narrowly tailored, well-guard-railed, lower-risk applications, it may be reasonable to permit a more limited premarket evidence package when coupled with rigorous post-market monitoring, real-world performance assessment, and appropriate safeguards to ensure continued safety and effectiveness.

At the same time, the success of such an approach depends on the existence of a clear and transparent postmarket oversight framework. The FDA has not yet fully articulated how ongoing monitoring of adaptive or GenAI-enabled systems would function in practice, including how performance drift, emergent risks, or degradation in clinical accuracy would be identified and addressed over time. Greater clarity regarding postmarket data collection, reporting expectations, and performance thresholds would help provide confidence that reduced premarket evidence can be responsibly balanced with robust lifecycle oversight.

Postmarket monitoring should extend beyond traditional adverse event reporting and include ongoing assessment of model performance across different clinical settings, patient populations, and user groups.¹⁹ Continuous reevaluation is particularly important for GenAI-enabled systems because performance characteristics and user interactions may change over time after deployment. Monitoring programs should be designed to detect performance drift, emerging sources of bias, unexpected behaviors, and other issues that may not be apparent during premarket testing.

For example, a GenAI-enabled system intended to assist with glioma classification or prediction of tumor characteristics may demonstrate strong performance during development and internal validation, particularly when trained on imaging and pathology datasets from a limited number of institutions.

¹⁸ Yu Huang et al., Deep Learning Achieves Neuroradiologist-Level Performance in Detecting Hydrocephalus Requiring Treatment, 35 J. Digital Imaging 1662 (2022), <https://doi.org/10.1007/s10278-022-00654-3>.

¹⁹ Ylenia Rotalinti et al., Identifying and Understanding Significant Change Due to Drift When Assessing AI Models in Healthcare: A Narrative Review, 2 BMJ Digital Health & AI e000085 (2026), <https://doi.org/10.1136/bmjdh-2024-000085>

However, model performance may be less reliable when applied to broader patient populations or different clinical environments unless the model has undergone rigorous external validation.²⁰ If the underlying dataset underrepresents patients treated in smaller community settings or lacks demographic diversity, the model's predictions may be less accurate when deployed more broadly. Because treatment planning and surgical decision-making may depend on these outputs, ensuring representative training data is critical to avoiding performance differences across patient populations.²¹

The FDA should also recognize that GenAI systems may reproduce historical treatment practices embedded in their training data, even when those practices do not reflect current standards of care. Clinical practice evolves over time as new evidence emerges, and GenAI models trained predominantly on historical data may reinforce outdated approaches or patterns of care rather than current standards. As a result, evaluation of GenAI-enabled devices should consider not only the representativeness of training data across patient populations, but also whether the underlying data reflect contemporary clinical practice. Transparency regarding the sources, timeframes, and characteristics of training data can help clinicians and regulators assess whether model outputs are likely to reflect current evidence-based care.

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 FDA's approach to lifecycle oversight for GenAI-enabled devices should recognize the important role that PCCPs can play in facilitating continuous improvement while maintaining safety and effectiveness. For GenAI-enabled technologies in particular, refinement following commercialization is often necessary to address evolving clinical workflows, improve usability, incorporate additional high-quality data, and strengthen performance across diverse patient populations. Requiring new premarket review for every modification could unnecessarily slow beneficial updates and discourage innovation without providing a public health benefit.²² The FDA should adopt a risk-based approach in which the level of re-benchmarking, validation, and regulatory oversight is proportionate to the nature of the change and its potential impact on device performance, intended use, and clinical decision-making.²³

PCCPs authorized as part of a device's premarket review can provide a structured and transparent framework for managing anticipated modifications within clearly defined boundaries. By specifying in advance the types of permissible changes, performance thresholds, and monitoring requirements, PCCPs can establish a baseline against which future updates may be assessed while ensuring that modifications remain consistent with the device's authorized intended use and risk profile. This

²⁰ Bartosz Szmyd et al., *AI-Driven Innovations in Neuroradiology and Neurosurgery: Scoping Review of Current Evidence and Future Directions*, 17 *Cancers* 2625 (2025), <https://pmc.ncbi.nlm.nih.gov/articles/PMC12384332/>

²¹ Gerard O'Sullivan et al., *Guiding Principles for the Responsible Development of Artificial Intelligence Tools for Healthcare*, 29 *Nature Med.* 1743 (2023), <https://pmc.ncbi.nlm.nih.gov/articles/PMC10066953/>

²² U.S. Food & Drug Admin., Health Canada & Medicines & Healthcare Products Regulatory Agency, *Predetermined Change Control Plans for Machine Learning-Enabled Medical Devices: Guiding Principles 2-3* (2021)

²³ U.S. Food & Drug Admin., *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions 8-10* (Aug. 2025)

approach allows the FDA to maintain oversight of meaningful changes while providing manufacturers with greater regulatory predictability and flexibility to implement lower-risk updates.

Certain categories of modifications may be well suited for inclusion within a PCCP because they are unlikely to affect the safety or effectiveness of a GenAI-enabled device when appropriately validated. Examples include refining prompts or system instructions, updating safety guardrails, expanding training datasets consistent with the authorized intended use, or improving outputs to reduce bias, hallucinations, or variability.^{24, 25}

By contrast, changes that could alter a device's intended use, level of autonomy, clinical role, target population, or output characteristics should generally remain subject to FDA review. Accordingly, the FDA should establish clear principles that describe when modifications can be implemented under an authorized PCCP, when a manufacturer's quality assurance procedures can appropriately manage them, and when additional premarket review is warranted. This would help ensure that oversight remains focused on changes that present meaningful risk while enabling timely implementation of lower-risk improvements that can enhance device performance, reliability, and patient care.

CONCLUSION

The AANS and CNS appreciate the opportunity to comment on the *Considerations for the Regulations of Generative AI-Enabled Medical Devices* discussion paper and look forward to continued engagement with FDA on this important issue. If additional information would be helpful or questions arise regarding these comments, please contact Natalie Torentinos, MPS, Manager of Regulatory Policy for the AANS/CNS Washington Committee, at ntorentinos@neurosurgery.org. Thank you for considering these recommendations and for your continued efforts to improve patient care.

Sincerely,



E. Antonio Chiocca, MD, PhD
President
American Association of Neurological Surgeons



Martina Stippler, MD
President
Congress of Neurological Surgeons

²⁴ U.S. Food & Drug Admin., Health Canada & Medicines & Healthcare Products Regulatory Agency, *Predetermined Change Control Plans for Machine Learning-Enabled Medical Devices: Guiding Principles* 2-4 (2021)

²⁵ U.S. Food & Drug Admin., *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan* 3-5 (2021).