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

August 2026

FDA U.S. FOOD & DRUG ADMINISTRATION

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

This paper is intended for discussion purposes only and does not represent draft or final guidance. This paper is not intended to propose or implement policy changes regarding how CDRH intends to regulate generative AI-enabled devices. This paper is not intended to communicate CDRH's proposed (or final) regulatory expectations, including its expectations for supporting evidence in future marketing submissions, but is instead meant to seek early input from groups and individuals outside the Agency and to advance a broader discussion among stakeholders on this topic. Finally, this paper is not intended to address whether the approaches discussed below are within FDA's existing legal authorities or whether new legal authorities would be necessary.

I. Introduction

Generative artificial intelligence (GenAI)-enabled medical devices hold transformative promise for patient care and the broader health ecosystem. At the same time, these devices may introduce unique risks when compared to traditional software and artificial intelligence (AI)-enabled medical devices.[1] FDA's Center for Devices and Radiological Health (CDRH) has previously identified certain regulatory challenges for GenAI-enabled medical devices[2] (hereafter referred to as "GenAI-enabled devices") and continues to contemplate ways to advance regulatory approaches for these devices. This discussion paper is intended to engage stakeholders on this topic and to solicit feedback on risk assessment, premarket evaluation, postmarket monitoring, and other topics relevant to the regulation of GenAI-enabled devices. CDRH's ultimate goal is to enable a nimble regulatory approach that employs least burdensome principles and allows patients timely access to safe and effective medical devices.

As with all medical devices, CDRH anticipates applying a risk-based approach to regulating GenAI-enabled devices, taking into consideration the product's intended use and technological characteristics. Furthermore, CDRH has long promoted a total product life cycle (TPLC) approach to the regulation of medical devices, including AI-enabled devices, and a TPLC approach is likely to be important to the regulation of GenAI-enabled devices.

[1] See FDA's AI-Enabled Medical Devices List available at: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices.

[2] See "Executive Summary for the Digital Health Advisory Committee Meeting: Total Product Lifecycle Considerations for Generative AI-Enabled Devices" available at https://www.fda.gov/media/182871/download.

II. Background

GenAI-enabled devices have unique characteristics and behaviors that are distinct from traditional software and AI-enabled devices that FDA regulates. They may accept open-ended inputs, perform multiple subtasks, and produce variable outputs to similar inputs. In some cases, they can also evolve over time through changes to features such as the underlying model, prompts, retrieval strategies, guardrails, orchestration logic, and/or user interface. Many are built on general-purpose foundation models developed by third-party entities, whose models offer varying levels of transparency into their training data, architecture, and evaluation methods, making it difficult to attribute specific behaviors and errors to the device or its underlying model. The generally broader and more flexible capabilities associated with GenAI-enabled devices may offer several benefits, including more personalized outputs, improved adaptability to novel inputs, enhanced user interaction, and support for more complex decision-making tasks. However, the same characteristics may introduce risks, including confabulations (or hallucinations) that may appear authentic to users, uncertainty in the bounds of a device's intended use, limited visibility into underlying third-party foundation models, and performance degradation of the device and its components across test environments and in real-world applications across the TPLC.

In November 2024, CDRH held a public advisory committee meeting of the Digital Health Advisory Committee (DHAC) to discuss TPLC considerations for GenAI-enabled devices.[3] In its Executive Summary for this DHAC meeting, CDRH discussed two broad categories of regulatory challenges applicable to GenAI-enabled devices: (1) challenges associated with applying a risk-based approach to classifying and determining regulatory requirements for GenAI-enabled devices; and (2) challenges associated with determining the types of valid scientific evidence ("evidence") for FDA's evaluation of the safety and effectiveness of such devices across the TPLC.[4]

In particular, CDRH discussed the possibility that new methodologies may need to be developed to evaluate the performance of GenAI-enabled devices. CDRH also acknowledged that, even with new methodologies, some GenAI-enabled devices may be challenging to review in the premarket setting because they have the potential to undergo continuous adjustments and, for devices with broad intended uses or those that use open-ended input and output formats, it may be unreasonable to evaluate every conceivable input that the device might encounter during deployment. Therefore, CDRH stated that it will be important for manufacturers to consider how to effectively evaluate and monitor the GenAI-enabled device, and the underlying foundation model as applicable, for its specific intended use in a way that can ensure that device accuracy, relevance, and reliability are maintained, once deployed. CDRH also stated that to ensure continued safety and effectiveness of GenAI-enabled devices, it may be

[3] For more information on this meeting, see November 20-21, 2024: Digital Health Advisory Committee Meeting Announcement - 11/20/2024.

[4] See "Executive Summary for the Digital Health Advisory Committee Meeting: Total Product Lifecycle Considerations for Generative AI-Enabled Devices" available at https://www.fda.gov/media/182871/download.

important for premarket evidence to be complemented by postmarket evidence gathered through robust device performance monitoring approaches.

The regulatory challenges applicable to GenAI-enabled devices and associated considerations are addressed in further detail in the Executive Summary for the November 2024 DHAC meeting. CDRH continues to be committed to working collaboratively with stakeholders to identify efficient, scientifically sound approaches for applying least burdensome principles to the evaluation of GenAI-enabled devices, informed by established best practices. CDRH's goal is to support innovation while promoting and protecting public health, and to establish regulatory approaches that will permit beneficial devices to reach clinicians, patients, consumers, and health systems when there are reasonable assurances of safety and effectiveness.

III. Definitions and Context

This discussion paper focuses on GenAI-enabled devices. To support efficient discussion, this section provides brief definitions of key terms drawing from FDA's Digital Health and Artificial Intelligence Glossary – Educational Resource and the context surrounding this discussion paper.

GenAI is a class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content.[5]

Foundation models are AI models trained using large, typically unlabeled datasets and significant computational resources that are applicable across a wide range of contexts, including some contexts that the models were not specifically developed and trained for (i.e., emergent capabilities). These models can serve as the basis upon which further models can be built and adapted for specific uses through further training (fine-tuning). These models can perform a range of general tasks, such as text synthesis, image manipulation, and audio generation. These models are based on deep learning architectures like transformers and can use either unimodal or multimodal input data.[6,7]

A **Large Language Model (LLM)** is a type of AI model trained on large text datasets to learn the relationships between words in natural language. These models can apply these learned patterns to predict and generate natural language responses to a wide range of inputs or prompts they receive to conduct tasks like translation, summarization, and question answering. These models are characterized by a vast number of model parameters (i.e., internal learned variables within a trained model).[8]

---

[5] See FDA Digital Health and Artificial Intelligence Glossary – Educational Resource.
[6] For a description of transformer architecture, see Vaswani A, et al. Attention Is All You Need. Advances in Neural Information Processing Systems. 2017:30. doi:10.48550/arXiv:1706.03762.
[7] See FDA Digital Health and Artificial Intelligence Glossary – Educational Resource.
[8] See FDA Digital Health and Artificial Intelligence Glossary – Educational Resource.

Agentic AI systems are GenAI-enabled systems that autonomously plan and execute multi-step tasks, use external tools, or take actions across a sequence of steps.

Multimodal is an approach for processing and integrating multiple different data types, aiming to capture and leverage the relationships between them for a better understanding of the input information or improved prediction performance. These data types may include text, images, audio, video, genomics, sensor data, etc. These different data types may be processed using a single multimodal network (e.g., based on neural network, or other architectures) or through separate unimodal networks (e.g., LLMs for text and convolutional neural networks for images) where the unimodal outputs are combined.[9]

As described in FDA guidance Multiple Function Device Products: Policy and Considerations, medical products may contain several functions. A **function** is a distinct purpose of the product, which could be the intended use or a subset of the intended use of the product. For example, a product with an intended use to analyze data has one function: analysis. A product with an intended use to store, transfer, and analyze data has three functions: (1) storage, (2) transfer, and (3) analysis.

FDA's current risk-based approach to the regulation of software functions applies to each software function of a product. As discussed in FDA guidance Policy for Device Software Functions and Mobile Medical Applications (or DSF-MMA guidance), many software functions are not medical devices, meaning such software functions do not meet the definition of "device" in section 201(h) of the FD&C Act (in some cases, because they are excluded from the definition of "device" pursuant to section 520(o) of the FD&C Act). FDA does not regulate such software functions as devices. Other software functions may meet the definition of a device, but because they pose a lower risk to the public, FDA may have communicated general enforcement discretion policies for these functions (meaning FDA has said that generally it does not intend to enforce requirements under the FD&C Act). In general, FDA applies its regulatory oversight to those software functions that meet the device definition and whose functionality could pose a risk to a patient's safety if the device were not to function as intended. Software functions that meet the definition of device are referred to as **device software functions**.

In this discussion paper, the term **GenAI-enabled device** is used to describe a product that contains one or more device software functions that are enabled by GenAI. Similarly, a **GenAI-enabled device software function** is a device software function that is enabled by GenAI. In some instances, a GenAI-enabled device software function may itself be a GenAI-enabled device when it is the only function in the device. However, a GenAI-enabled device can also contain multiple device functions, which may not all be GenAI-enabled, and/or "other functions"[10] that are not the focus of FDA's regulatory

[9] See FDA Digital Health and Artificial Intelligence Glossary – Educational Resource.

[10] For purposes of this discussion paper, "other functions" are functions that do not meet the definition of device; meet the definition of device but are exempt from premarket notification (i.e., 510(k)-exempt); or meet the definition of device but for which CDRH has expressed its intention not to enforce compliance with FDA requirements. For more information, see FDA guidance Multiple Function Device Products: Policy and Considerations.

oversight. CDRH acknowledges that a GenAI-enabled device software function can be integrated into a broader device, with other hardware and software components; this discussion paper addresses the GenAI-enabled component of that device, but not the entire device.

It is important to note that FDA does not regulate GenAI as such; it regulates medical devices, including GenAI-enabled devices. This is consistent with how FDA considers other types of technology in its regulation of devices: FDA does not regulate software as such, hardware as such, or AI as such, but regulates products that meet the definition of device under the FD&C Act, applying risk-based oversight based on the regulatory controls needed to provide a reasonable assurance of safety and effectiveness.

IV. Considerations for the Assessment of Risk for GenAI-Enabled Devices

Assessment of risk is critical to risk-proportionate regulation of GenAI-enabled devices. Risk assessment may affect FDA's scope of regulatory oversight, its evidentiary requirements in the premarket setting, and its expectations surrounding postmarket monitoring and change control.

To support consistent thinking about risk for GenAI-enabled software functions, a two-axis framework is a possible organizing heuristic **(Figure 1)**. Such a framework draws conceptual structure from analogous matrices used in other FDA guidance documents.[11] The framework places device activity on one axis and the "consequences," or severity of harm of relying on an incorrect device output, on the other. It implies a gradient of increasing risk extending from the lower-left corner to the upper-right.

[11] See, for example, FDA draft guidance Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products and FDA guidance Assessing the Credibility of Computational Modeling and Simulation in Medical Device Submissions.

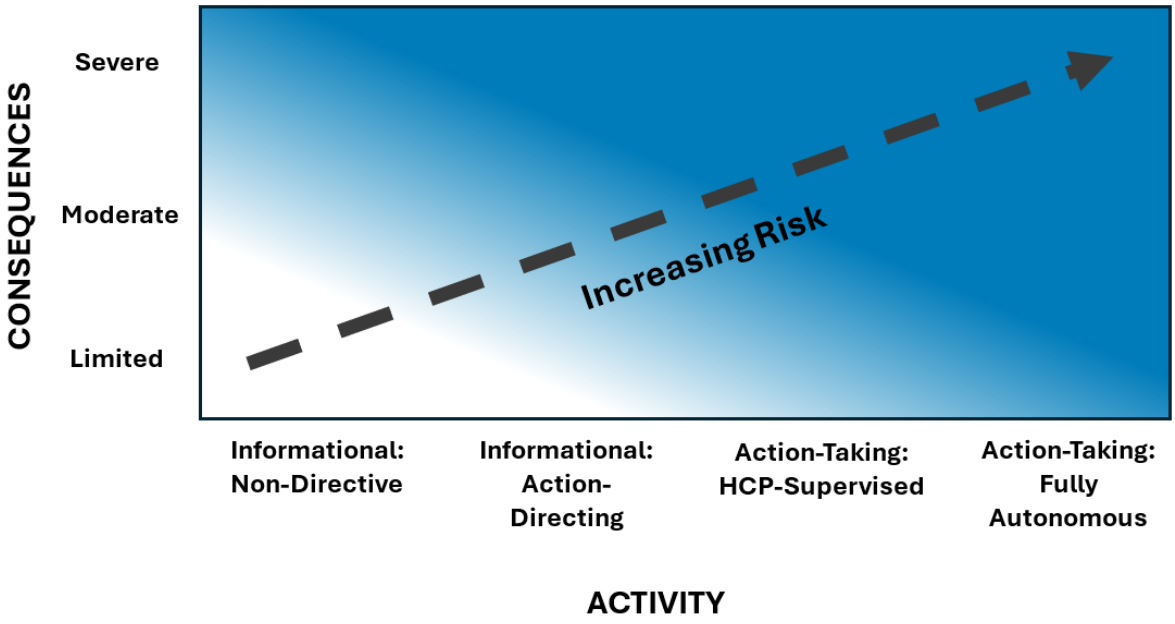

Figure 1. A possible two-axis framework for thinking about risk for GenAI-enabled software functions. The horizontal axis represents the activity performed by the function, increasing in the degree and independence of device activity from left to right. The vertical axis represents the consequence of relying upon an incorrect output. Risk increases from the lower left toward the upper right. HCP = healthcare professional.

The activity axis recognizes that the risk of a software function depends in part on how independently it directs or takes action. A function that provides non-directive information (for example, a risk score for a future cardiovascular event) is meaningfully different from one that, while nominally informational, directs an action (for example, a strongly worded patient-facing recommendation about whether to seek emergency care)[12]. A function that acts with continuous healthcare professional (HCP) supervision is meaningfully different from a function that acts in fully autonomous fashion.

The consequences axis recognizes that the risk of a software function also depends on the severity of harm that may result from using and relying upon an incorrect output. An incorrect recommendation about an over-the-counter remedy for a minor symptom is meaningfully different from an incorrect recommendation about adjusting an insulin dose.

CDRH seeks stakeholder feedback on possible use of this two-axis framework as a heuristic for thinking about risk for GenAI-enabled devices. It also seeks feedback on the following additional considerations for the assessment of risk as related to the two-axis framework:

[12] These and other examples in this section are illustrative only. They are meant to demonstrate concepts of the two-axis framework and are not intended to mean that the software discussed is necessarily a device or is a software function that is the focus of FDA's regulatory oversight.

"Action-directing" functions. As noted above, certain GenAI-enabled informational functions direct action for a patient or HCP. CDRH is considering treating these "action-directing" functions as higher risk functions than GenAI-enabled functions that provide simple non-directive information.

CDRH notes that the distinction between "non-directive" and "action-directing" information may not be binary but rather would likely involve a continuum. The degree to which informational functions direct users toward a particular action might vary significantly. For example, an informational function may provide general information ("lisinopril dosages are sometimes increased when blood pressure remains above the treatment goal"), connect general practice to the user's circumstances ("in situations similar to this, clinicians often increase the lisinopril dosage"), endorse a specific action ("I recommend increasing the lisinopril dosage"), or provide a specific instruction ("increase the lisinopril from 10 mg to 20 mg daily"). CDRH is considering whether the degree to which an informational function may be directive may depend on the substance and context of the output, not solely on whether it uses words such as "recommend," "should," or "consider." CDRH is also considering whether a patient-facing informational function may not become any less directive because it includes a "talk to your doctor" or an "I am not a medical professional" statement in addition to the "action-directing" information.

When a function is considered "action-directing," CDRH is considering whether its presumptive risk may be moderated by its location on the consequences axis. For example, in the case of advisory information on self-treatment, a GenAI-enabled function directing a patient to apply over-the-counter hydrocortisone cream for poison ivy may be lower risk than a GenAI-enabled function directing a patient to increase or decrease their basal insulin dose. In the case of care triage, a GenAI-enabled function that advises a parent on whether or not to bring their child with a sore throat to a pediatrician or urgent care facility may be lower risk than a GenAI-enabled function that directs a patient to seek or not seek emergency department care for their chest pain.

"Action-taking" functions. CDRH is considering that a GenAI-enabled "action-taking" function, such as the affirmative assignment of a clinical diagnosis, the prescription of a medication, or the initiation of a clinical order set, is typically higher risk than a GenAI-enabled informational function. However, the presumptive risk associated with a GenAI-enabled action-taking function is also moderated by its location on the consequences axis. For example, the severity of harm that may result from the autonomous prescription of antibiotics for a confirmed strep throat infection may be substantially lower than the autonomous initiation of a thrombolytic therapy order set in a stroke workflow, even given similar levels of HCP supervision.

Measurement and signal processing functions. GenAI-enabled *in vitro* diagnostic, measurement, and signal processing functions present special considerations around risk assessment because they often produce outputs that are not directly assessable by the user, and the user may not be able to identify and avoid relying upon incorrect outputs. CDRH is considering that although these functions yield non-directive information, they may nonetheless be of higher risk because the user typically cannot

independently evaluate the basis for the output (and therefore the function may migrate higher on the consequences axis).[13]

Patient-facing informational functions versus HCP-facing informational functions. GenAI-enabled patient-facing informational functions may present different considerations from GenAI-enabled HCP-facing informational functions. HCPs may be better equipped than patients to interpret an output in context, to recognize its limitations, and to integrate it with other clinical information such that they can identify an incorrect output and not rely upon it; patients, lacking this training, may not be able to identify an incorrect output, may rely upon it, and may take downstream actions that adversely affect outcomes, including delaying needed care or pursuing inappropriate self-treatment. Such considerations could be captured within the two-axis risk framework by shifting certain functions higher on the consequences axis when they are patient-facing rather than HCP-facing. On the other hand, there may be important benefits associated with the democratization of clinical information, with patient empowerment and engagement, and with avoiding medical parentalism that may underestimate patient capability. CDRH seeks stakeholder input on whether and when the delivery of clinical information to users who lack the domain knowledge needed to independently evaluate the output should be considered higher risk.

Generalist HCP-facing versus specialist HCP-facing functions. A closely related question arises with respect to GenAI-enabled functions intended for use by HCPs without specialty-level training in the clinical domain the function is intended for. A function that helps a generalist HCP approach a question typically handled by a specialist may meaningfully extend access to specialty knowledge, while a function whose safe use depends on contextualization by specialist expertise may present elevated risk when used by HCPs lacking that expertise to independently evaluate outputs. CDRH seeks input on whether and how this distinction might inform the assessment of risk.

Beyond the two-axis risk framework, CDRH also seeks feedback on the following additional considerations relevant to the assessment of risk for GenAI-enabled devices:

Multi-turn conversations. Conversational GenAI-enabled devices may produce outputs in the context of an extended exchange. A device may begin by performing an informational function and migrate, over the course of a conversation, to an action-directing function. CDRH is considering an approach for assessing risk of such devices that would consider the device's behavior across realistic conversational trajectories, not only at the level of individual functions.

Care escalation functions. CDRH notes that for GenAI-enabled functions involving care escalation decisions, such as advising whether to seek emergency care, both under-escalation and over-escalation can impact assessment of risk of such functions. Under-escalation, in which a function fails to direct a patient toward needed care, can

[13] This consideration seems consistent with the independent-review criterion of the clinical decision support exclusion under section 520(o)(1)(E).

result in delayed treatment with potentially serious consequences. Over-escalation, in which a function systematically directs patients toward emergency care that is not clinically warranted, can produce harms that include patient anxiety, unnecessary tests or procedures, unnecessary emergency department utilization, increased resource burden on healthcare systems, and erosion of patient trust in ways that may reduce appropriate care-seeking over time. CDRH is considering whether assessment of risk in GenAI-enabled escalation functions should consider both directions of error.

A. Discussion questions

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.
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?
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?
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?
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?

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?

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

CDRH recognizes that evaluation approaches developed for software with bounded inputs and fixed outputs may not be appropriate for GenAI-enabled devices. Such evaluation has traditionally relied on extensive testing across a representative sample of device inputs and outputs; for GenAI-enabled devices, the range of possible inputs and outputs may be too large for such testing to be practical. As described in Section II, CDRH has previously recognized that performance evaluation of GenAI-enabled devices and their functions may require new methodologies.

Recognizing the potential need for new evaluation approaches for GenAI-enabled devices, several authors have advocated for models inspired by the evaluation and credentialing of human clinicians. Patel and Blumenthal[14] argue that GenAI is too broad, adaptive, and opaque to be evaluated effectively through current medical device regulatory frameworks and instead propose competency-based regulation modeled on medical training, licensure examinations, supervised practice, periodic reevaluation, and public reporting. Bergman, Wachter, and Emanuel[15] extend the concept to autonomous GenAI through a licensure-like framework with defined scope of practice, time-limited certification, ongoing monitoring, and explicit accountability for developers and deploying institutions. Freyer et al.[16] acknowledge the compelling aspects of a human clinician-inspired regulatory approach but raise questions about accountability, noting that human clinicians face professional, legal, and reputational consequences when expected standards are not met, and arguing that analogous mechanisms are needed for flawed GenAI-enabled devices.

CDRH is considering a competency-based approach to premarket evaluation of GenAI-enabled devices that is also inspired, at a high level, by how human clinicians are evaluated and credentialed, but would be adapted for the technical, practical, and legal considerations applicable to the regulation of medical devices. Clinicians, as well as GenAI-enabled devices, can ingest broad and varied information and can produce open-ended responses across a wide range of situations. Moreover, clinicians are not

[14] Patel B, Blumenthal D. A Novel Approach to Overseeing the Clinical Application of Generative AI. JAMA Health Forum. 2026;7(3):e256947. doi:10.1001/jamahealthforum.2025.6947.
[15] Bergman A, Wachter RM, Emanuel EJ. A Licensure Framework for Autonomous Clinical AI. JAMA. 2026;335(20):1751-1754. doi:10.1001/jama.2026.5483.
[16] Freyer O, Jayabalan S, et al. Overcoming Regulatory Barriers to the Implementation of AI Agents in Healthcare. Nature Medicine. 2025;31(10):3239-3243. doi:10.1038/s41591-025-03841-1.

evaluated through exhaustive testing of every scenario they may encounter; instead, they are evaluated through a combination of structured assessments of underlying knowledge and reasoning (including medical licensing, board examinations, and specialty board certifications) and supervised practice in real clinical environments with progressively greater independence (including during medical school clerkships, postgraduate residency training, and fellowships), followed by ongoing assessment after independent practice begins. Below, CDRH outlines and seeks input on a possible approach to the premarket evaluation of GenAI-enabled devices inspired by these considerations.

A. A Competency-Based Approach

CDRH is considering a competency-based approach to premarket evaluation of GenAI-enabled devices, consisting of non-clinical device benchmarking (Section V.B) and clinical confirmation (Section V.C). Under such an approach, the final user-facing device, as configured and intended to be deployed for real-world use—and not the foundation model standing alone or other isolated subcomponent—would be evaluated.

Importantly, a competency-based approach could be tailored to the device's intended use and be proportionate to its risk, potentially utilizing the two-axis risk framework described in Section IV. For example, the device activity and the consequences of relying on an incorrect output could help inform the nature, rigor, and amount of evidence needed, including the elements evaluated through benchmarking, the range and difficulty of test conditions, and the extent of clinical confirmation.

The sections below describe high-level considerations for how a competency-based assessment of a GenAI-enabled device might be carried out. These descriptions are intended only to stimulate discussion and obtain stakeholder feedback. CDRH is committed to working collaboratively with stakeholders across government, industry, and academia to identify efficient, scientifically sound approaches to applying least burdensome principles to the evaluation of GenAI-enabled devices.

B. Device Benchmarking

Device benchmarking[17] is conceived as evaluating whether the device, in its deployed or representative configuration, demonstrates the clinical knowledge, analytic capabilities, safety behavior, communication, and generalizability to support reasonable assurance of safety and effectiveness of the device for its intended use. It is conceived as a scalable, high-throughput approach to non-clinical evaluation, building on testing methodologies that can characterize device capabilities across a broad range of inputs.

[17] CDRH recognizes that industry stakeholders may use "benchmarking," "evaluation," or other terms—sometimes interchangeably and sometimes with slight distinctions—in reference to the methods described in this section. CDRH uses the "benchmarking" term throughout to avoid confusion with the term "evaluation," which can imply other types of testing for medical devices generally.

CDRH notes that the term “benchmarking” is broadly understood by industry stakeholders to mean the use of well-defined, published, reusable tests and datasets (“assets”) to measure and compare the performance of AI-enabled products at a specific point in time. This is consistent with how this term is used in this discussion paper, although CDRH is considering that sponsors might not only use well-defined tests and datasets developed by third-party entities, but also propose their own specialized tests and datasets tailored to their specific device and its functions. CDRH recognizes that there may be challenges with publicly available benchmarking assets due to data contamination, saturation, and lack of representativeness of real-world conditions[18] and is interested in feedback from stakeholders on how this may impact the approaches described in this paper.

1. Benchmarking Elements

CDRH is considering benchmarking to be composed of several elements. CDRH does not anticipate that all elements would necessarily apply to all devices; rather, elements would be chosen based on applicability to the device’s intended use and risk profile. CDRH is considering the following benchmarking elements:

- **Safety**:
  - S.1 Safety-critical recognition and escalation;
  - S.2 Scope maintenance and boundary adherence; and
  - S.3 Calibration, uncertainty communication, and clinical deferral
- **Clinical Proficiency**:
  - E.1 Clinical knowledge and task fidelity;
  - E.2 Information gathering and clinical analysis;
  - E.3 Quantitative and measurement analysis; and
  - E.4 Communication quality and user comprehension
- **Generalizability**:
  - R.1 Robustness, reliability, and reproducibility; and
  - R.2 Subgroup performance
- **Agentic AI Capabilities**:
  - A.1 Additional competencies applicable only to agentic GenAI-enabled devices

[18] Garcia V, Sidulova M, Badano A. Performance Assessment Strategies for Language Model Applications in Healthcare. Artificial Intelligence in the Life Sciences. 2026:9. doi:10.1016/j.ailsci.2026.100162.

Appendix A describes in greater detail how CDRH is considering each element.

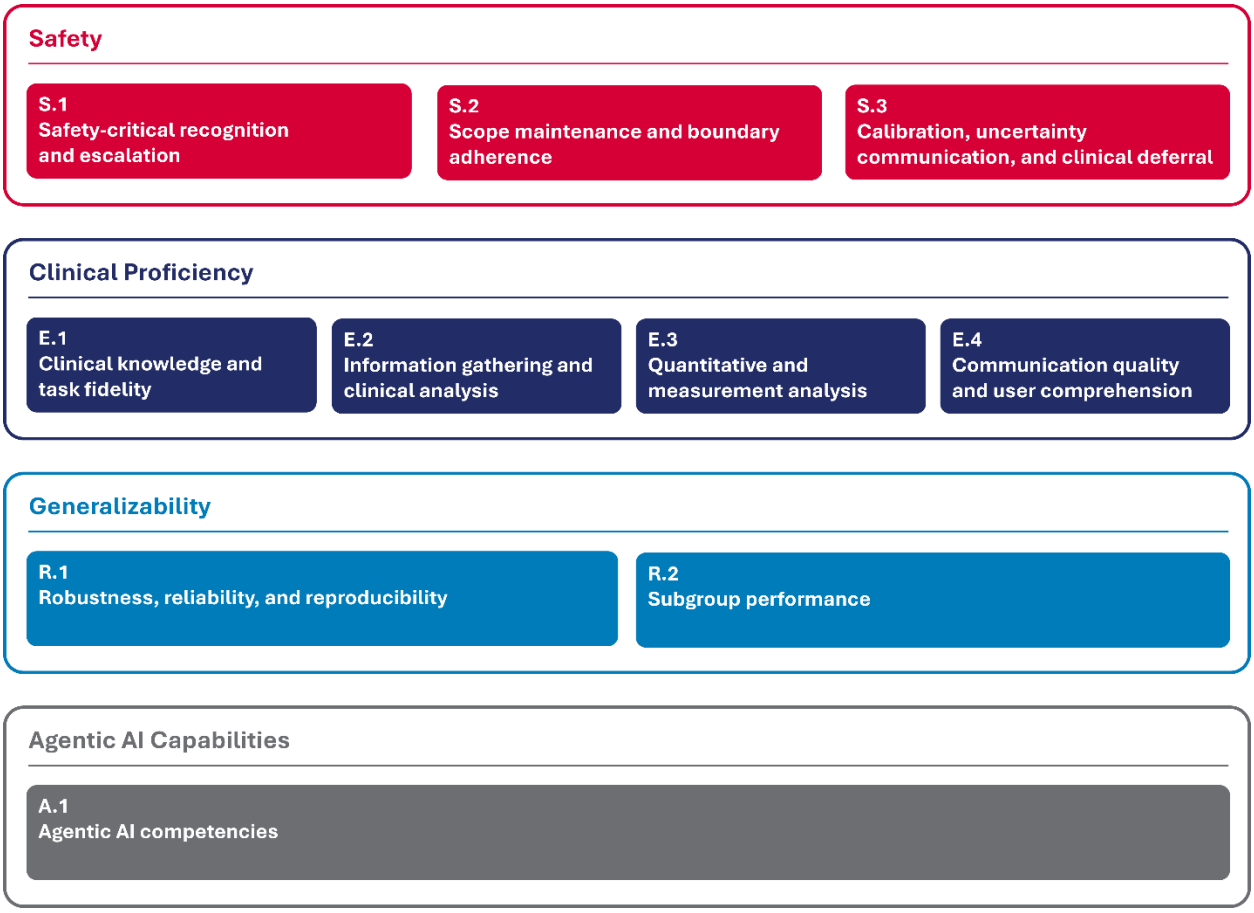

Figure 2. A possible approach to device benchmarking. Elements would be chosen based on applicability to the device's intended use and risk profile.

2. Methodological Approaches and General Principles

CDRH envisions that for benchmarking conducted under a competency-based approach, sponsors would define the scope of testing based on the device's intended use and risk profile, prespecify evaluation methods, describe and justify scoring rubrics, identify qualified expert adjudicators where applicable, and prespecify and justify acceptance criteria. CDRH recognizes that there could be opportunities for sponsors to leverage methodologies such as synthetic data generation and simulation tools, including virtual patient avatars, to conduct testing in a manner that is predictable, standardized, and amenable to automation. Such infrastructure could potentially support not only initial evaluation but also ongoing assessment of device performance over time and in response to modifications; in particular, as discussed in Section VI, the initial device benchmarking may serve as a reference against which later changes are evaluated.

CDRH is considering that several general principles may apply across benchmarking, such as:

- Consistent with expectations for non-clinical performance testing across devices,[19] test methods and acceptance criteria would be prespecified prior to testing.
- Where device outputs are evaluated against scoring rubrics, such rubrics would be grounded in established clinical guidelines or validated through qualified expert input and fit-for-use within the assessment context, including with respect to the specific patient population, clinical setting, and intended use of the device.
- Where expert adjudication is warranted, it would include adjudicators that are structurally independent from the device sponsor and, where the device incorporates a third-party model, from the developer of that model, and possess qualifications aligned with the clinical domain and complexity of the outputs under review. These considerations would still be applicable when the expert adjudicator is itself an LLM.
- Benchmarking would be expected to be of sufficient clinical and functional scope to evaluate the device for its intended use. For example, a GenAI-enabled device intended to prescribe medications for cardiovascular conditions may prescribe medications that can affect comorbid diseases in other clinical domains or interact with medications prescribed for non-cardiovascular conditions; in such cases, it may be appropriate to expect benchmark testing to assess clinical knowledge and analysis across other relevant clinical domains. Similarly, because a GenAI-enabled device may perform differently depending on its deployment environment, it may be appropriate to expect benchmark datasets to be representative of the deployment contexts in which the device is expected to be used.

C. Clinical Confirmation

CDRH anticipates that under a competency-based approach, device benchmarking, as described above and however thorough, may not fully establish how a GenAI-enabled device will perform in real clinical use. GenAI-enabled devices interact with users, workflows, and patient populations in ways that may not be captured by structured benchmarking. CDRH is therefore considering that clinical confirmation may be a necessary step in addition to benchmarking to determine whether a GenAI-enabled device performs as intended for its intended population under real-world or clinically representative conditions.

Importantly, CDRH is considering that under such an approach, clinical confirmation might not require a prospective clinical study in every case. Rather, sponsors would be expected to provide evidence that the device performs in real or clinically representative use, with the type and rigor of evidence consistent with the device's intended use and

[19] See, e.g., FDA guidance Recommended Content and Format of Non-Clinical Bench Performance Testing Information in Premarket Submissions.

proportionate to its risk profile, potentially utilizing the two-axis risk framework described in Section IV. The appropriate approach may also depend on the type of patient data the device acts on.

1. Clinical Confirmation Approaches

A range of possible approaches might be considered to generate evidence of device performance in real or clinically representative use.[20] These might be used individually, in sequence, or in combination. In approximate order of increasing rigor and patient exposure, these may include:

- **Retrospective evaluation on real patient inputs.** The device is applied to previously collected patient inputs, with outputs compared to an established reference standard or to expert clinical adjudication. Expert clinical adjudication might be established through multi-clinician consensus, with adjudicator independence from the device manufacturer and foundation model developer (as applicable) documented. Where real patient inputs are limited, synthetically generated inputs modeled on real patient data might, with appropriate justification, supplement this evaluation.
- **Shadow deployment.** The device operates in a live clinical workflow on real patients, but its outputs are not shown to clinicians or patients and do not affect care. Outputs are recorded and compared against decisions actually made and, where appropriate, downstream outcomes.
- **Standardized patient interactions.** Trained patient actors or representative individuals who could plausibly present as patients are directed to follow defined scenarios, with outputs evaluated against prospectively defined criteria. Depending on risk profile, this might serve as a proxy for performance with actual patients.
- **Clinician adjudication of real cases.** Qualified, independent clinicians review samples of real patient inputs and device outputs against prospectively defined criteria. Depending on the device, this might take the form of blinded review, in which a clinician reviews real patient inputs without seeing the device's outputs and independently determines the appropriate response for comparison against the device output, or unblinded review of full interaction records to assess the quality and appropriateness of outputs.
- **Prospective clinical study.** Depending on the device, a prospective clinical study, and in some instances a randomized controlled clinical trial, might be the appropriate confirmatory method.

CDRH recognizes these approaches may incorporate real-world data (RWD) and/or data collected outside the United States (OUS) and seeks input on specific considerations for the use of such data in the clinical confirmation of GenAI-enabled devices.

[20] Certain regulatory requirements, including but not limited to, requirements in 21 CFR Parts 50, 56 and 812, may apply to these approaches. However, the extent of their application is not in scope of this discussion paper.

Across these approaches, relevant considerations might include whether the device performs consistently across clinically relevant subgroups within the intended use population. Where subgroup evaluation would be difficult, particularly for small or hard-to-sample subgroups where real-world data is limited, CDRH is considering that synthetically generated data modeled on real patient data could potentially supplement evaluation, assuming such synthetic data is representative of the environment in which the device is intended to be deployed.

CDRH acknowledges that the approaches outlined above may not be powered around traditional effectiveness endpoints, and is interested in stakeholder input on how endpoints and sample sizes might be determined.

D. Other Considerations for Premarket Evaluation of GenAI-Enabled Devices

1. Standards of Performance

A recurring question for premarket evaluation of GenAI-enabled devices is the standard against which performance should be measured and how adequate performance should be defined. Conventional performance evaluation often compares a device output against a known correct output. The outputs of a GenAI-enabled device, however, may be open-ended, such that a single correct response often does not exist and many different responses may be acceptable.

CDRH is considering that for many GenAI-enabled devices, performance might be compared to that of a panel of qualified clinicians whose consensus reflects the applicable standard of care, or to that of a median clinician in practice. With such an approach, the relevant question might be whether the device performs as well as, or better than, a reference clinician on the same task, while recognizing that additional questions may arise regarding whether generalist or specialist physicians should be used for comparison. For other GenAI-enabled devices, a more traditional ground truth or clinical reference standard may be available and appropriate, e.g., a diagnosis confirmed by biopsy or by adjudicated expert consensus. Considerations for selecting the comparator might include how the device is actually used, including the combined performance of the clinician and device working together as a human-AI team versus the device working in a fully autonomous workflow, depending on the intended use.

CDRH seeks input on considerations in selecting comparators, performance comparisons to generalist versus specialist physicians, how both human-AI team and AI-working-alone performance might be evaluated, and ways to define performance thresholds and acceptance criteria.

2. Potential Roles for Independent Third Parties

Another consideration is whether qualified, independent third parties might play a role in benchmarking and clinical confirmation, for example, by providing larger

or sequestered evaluation datasets, offering independence between development and testing, or serving as independent expert adjudicators.

CDRH notes that existing programs already provide mechanisms for leveraging qualified third parties in the evaluation of medical devices, and CDRH is considering whether analogous third-party frameworks could be utilized to support the assessment of GenAI-enabled devices. For example, the Accreditation Scheme for Conformity Assessment (ASCA) Program is an initiative under which ASCA-recognized accreditation bodies accredit laboratories that can perform testing on devices in accordance with FDA-recognized voluntary consensus standards and test methods. FDA's Medical Device Development Tool (MDDT) program is a way for FDA to qualify tools that medical device sponsors can choose to use in the development and evaluation of medical devices, helping to show how medical devices work in terms of safety, effectiveness, and other aspects of performance.

CDRH seeks input on what role(s), if any, independent third parties might play in the premarket evaluation approach described above. For example, such roles might include:

- Conducting or evaluating the competency-based assessment of a GenAI-enabled device (or portions of the competency assessment) and issuing a determination (or a certificate) that CDRH could consider in its review of the device;
- Maintaining sequestered evaluation datasets and benchmarks that are developed independently from the device manufacturer;
- Serving as independent expert adjudicators;
- Developing tools qualified by FDA to contribute to device development and regulatory review; or
- Developing standards that might be recognized by FDA to promote consistency across testing and review.

E. Discussion questions

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?
8. How might the two-axis risk framework described in Section IV be considered within the competency-based approach described above to help determine the level of evidence needed for a premarket submission?
9. Would a benchmarking structure such as the one described above be likely to provide adequate evidence of clinical knowledge, analytic capabilities, safety behavior, communication, and generalizability to support a reasonable assurance of safety and effectiveness? Are there elements that are missing,

redundant, or inappropriately categorized? Are there externally developed standards that could be leveraged?

10. CDRH recognizes that publicly available benchmarking assets may be subject to data contamination, saturation, and limited real-world representativeness. How should a sponsor establish that performance on a given benchmark predicts safe and effective real-world behavior for the device's intended use? What evidence should support the construct validity of a benchmark used to gate device evaluation, and what role should sponsor-developed benchmarks play given potential concerns around independence and optimization to the test?
11. CDRH is considering that clinical confirmation for a GenAI-enabled device might not require a prospective clinical study in every case, and has described above a range of approaches of increasing rigor and patient exposure. How might a sponsor select and justify a confirmation approach tailored to a device's intended use and proportionate to the device's risk profile? Are there device types or risk profiles for which one or more of these approaches would be insufficient or inappropriate? How might the anticipated distribution of real-world inputs be taken into consideration? Are there other methods of clinical confirmation that might help inform the evaluation of GenAI-enabled devices?
12. How should sponsors achieve statistically meaningful performance measurement for GenAI-enabled devices? Where synthetically generated inputs supplement real patient data, how should sponsors account for differences between the synthetic and real-world distributions when estimating performance, and under what conditions, if any, is it appropriate to combine benchmarking evidence and clinical confirmation evidence to support a single performance estimate?
13. For which clinical domains, device functions, or subpopulations is synthetic data particularly well-suited, or particularly inadequate, as a supplement to real-world evidence? What safeguards would mitigate the risk that synthetic data generated by models of the same class as the device under evaluation reproduces the very performance gaps the evaluation is intended to detect, particularly for underrepresented subgroups?
14. For open-ended device outputs, how should performance comparators and acceptance criteria be selected? When a panel of qualified clinicians serves as the comparator, how should the applicable standard (for example, the standard of care versus the performance of a median clinician in practice) be defined and justified? Should generalist or specialist physicians be used as a performance standard? When should human-AI team performance, rather than the device operating alone, serve as the basis for evaluation?
15. Are there ways in which performance might be assessed relative to the care, technology, or course of action likely to occur in the absence of the device, rather than to the comparators described in this section? What approaches might be used to identify and justify a comparator such as unaided clinical judgment, delayed specialist review, or no intervention?
16. Should independent third parties be involved in some or all aspects of a competency-based approach, including device benchmarking and clinical confirmation? If so, in what ways might qualified, independent third-party

participation contribute to this approach, and what qualifications and independence criteria should apply? Are there aspects of the assessment for which third-party involvement would be impractical or inadvisable? What safeguards or program design features would be critical to prevent third-party participation from limiting competition or preventing innovation?

17. CDRH is exploring whether a competency-based approach could be applied to devices that incorporate a variety of underlying model architectures, such as multimodal vision-language models and generative or predictive world models. Are there aspects of the described approach that might be ineffective or inapplicable to these devices? What additional or distinct considerations might need to be addressed?

VI. Postmarket Monitoring for GenAI-Enabled Devices

As discussed in Section II, the novel capabilities of GenAI-enabled devices present unique challenges for premarket evaluation. Specifically, these devices produce varied, open-ended outputs in response to similar inputs and may undergo continuous adjustments following deployment, making it difficult for premarket testing to fully capture performance. The need for new postmarket evidence approaches was highlighted in the Executive Summary for FDA's 2024 DHAC Meeting, where CDRH emphasized that it is important for manufacturers to consider how to effectively evaluate and monitor GenAI-enabled devices once deployed. CDRH is considering whether it is appropriate to accept greater premarket uncertainty regarding a GenAI-enabled device's benefit-risk profile through greater reliance on postmarket monitoring.[21]

A. Postmarket Monitoring Approaches

CDRH recognizes that there may be a variety of different approaches to postmarket monitoring of GenAI-enabled devices, and that the scope, frequency, and level of evidence of such approaches might vary in proportion to the risk profile of the device. Examples of possible approaches include:

- **Periodic device benchmarking:** The sponsor reassesses the device against the prespecified benchmarking thresholds described in Section V.B, on a defined cadence and after defined triggering events, including changes to the underlying model or other components of the deployment architecture.
- **Periodic sample-based clinician review:** Qualified, independent clinician adjudicators review samples of real-world inputs and outputs against prospectively defined criteria, sampled to reflect the range of clinically relevant presentations and interaction patterns the device encounters in real-world use.

[21] For more information, please see FDA guidance Consideration of Uncertainty in Making Benefit-Risk Determinations in Medical Device Premarket Approvals, De Novo Classifications, and Humanitarian Device Exemptions.

- **Performance degradation monitoring:** The sponsor monitors for performance degradation (e.g., drift) that may result from changes in the input population, the data environment, or underlying model components, specifying the analyses and thresholds to detect such degradation.

CDRH seeks feedback on these approaches, as well as the feasibility and practicality of implementing these or other approaches to postmarket monitoring for GenAI-enabled devices. CDRH also seeks feedback on whether some aspects of postmarket monitoring can be enabled by machine-based supervisory agents, and is interested in stakeholder feedback on considerations for such an approach.

B. Shared Ecosystem Responsibility

While manufacturers are responsible for the postmarket monitoring of their devices, CDRH recognizes that the broader ecosystem within which GenAI-enabled devices will be deployed includes clinicians, patients, healthcare institutions, payers, professional societies, public-private consortia, standards-setting bodies, and other federal and state authorities, each with potential roles to play in the deployment, monitoring, reporting, and ongoing evaluation of GenAI-enabled devices. CDRH seeks to understand what roles and responsibilities these stakeholders may play for various approaches to postmarket monitoring.

C. Considerations for Postmarket Modifications

CDRH understands that GenAI-enabled devices may undergo changes after market authorization and once the device is deployed. Some changes may be intentional and discrete modifications actively implemented by the sponsor, such as a software update, algorithm revision, retraining event, or change to the device's intended functionality. Other changes may be better understood as model-evolution changes, i.e., changes that occur more passively, continuously, or incrementally because the device was designed to learn from new data or adapt during use without a distinct sponsor-initiated update. Still other changes may be unplanned changes arising from updates to an underlying third-party foundation model. A change of any of these types may have the potential to affect the safety and effectiveness of the device.

CDRH is considering several possible approaches to the oversight of postmarket changes to GenAI-enabled devices, with requirements that may range from documentation in a sponsor's quality management system to FDA authorization before changes are implemented. A Predetermined Change Control Plan (PCCP)[22] could be one mechanism to facilitate certain device changes without requiring a new premarket submission.

CDRH anticipates that a manufacturer's premarket competency-based assessment, such as the one described in Section V, could potentially form the baseline for ongoing

[22] For more information, see CDRH's Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions.

assessment when changes to a marketed GenAI-enabled device are made. A GenAI-enabled device that has been benchmarked against a defined set of capabilities for premarket authorization could potentially be “re-benchmarked” against the same capabilities following a modification, providing evidence to determine whether the modification has affected safety or effectiveness of the device. Such an approach would underscore the need to carefully design a competency-based assessment in the premarket setting.

Devices built on third-party foundation models raise particular change-control considerations because changes to the underlying model may be initiated by the foundation model developer rather than the device manufacturer. The use of these third-party foundation models may increase over time, and CDRH’s regulatory approach may need to account for this. Section VII.A seeks input on a possible approach to these considerations in more detail.

D. Discussion questions

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?
19. Please comment on the potential approaches to postmarket performance evaluation, including periodic re-benchmarking, sample-based clinician review, and performance degradation monitoring. What additional approaches should CDRH consider, and how should the cadence and triggering events for reassessment be determined?
20. Please comment on whether the proposed approaches to postmarket monitoring can be facilitated by machine-based supervisory agents. What considerations, including the evaluation and reliability of the supervisory agent itself, should CDRH take into account for such an approach?
21. What roles might clinicians, healthcare institutions, professional societies, standards-setting bodies, and other stakeholders appropriately play in postmarket monitoring, and how can these roles be structured without diffusing manufacturer accountability?
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?

23. How might PCCP concepts or other change-control approaches be adapted for GenAI-enabled devices when the nature or scope of future modifications cannot be fully prespecified?
24. For devices built on third-party foundation models, changes to the underlying model may be initiated by the third-party model developer rather than the device manufacturer. How can a manufacturer detect, evaluate, and respond to such changes in a timely manner? What mechanisms—for example, contractual, technical, or through a PCCP—could provide reasonable assurance that third-party developer-initiated changes do not compromise the safety or effectiveness of the device?

VII. Other Topics

A. Foundation Model Device Master Files (MAF)

CDRH recognizes that GenAI-enabled devices may be built on or incorporate third-party foundation models that may control model refusal behavior, content policies, output formatting, version control, and safety-critical behaviors integral to device safety and effectiveness. Given the potential contribution of these foundation models to overall device behavior, CDRH seeks feedback on the feasibility of voluntary Foundation Model MAFs, leveraging the existing Device Master File program,[23] under which foundation model developers and platform providers could voluntarily submit information such as structured model cards or system cards to FDA.[24] As with all MAFs, these files would be held by FDA confidentially and could be referenced by sponsors, with the file holder's authorization, in support of individual premarket submissions. The information would be made available to premarket review teams on a reference basis, enabling more consistent and informed review of GenAI-enabled devices built on a given model. Under such an approach, submission of a Foundation Model MAF would not constitute authorization of the underlying model for any device intended use, and sponsors would remain responsible for independently demonstrating the safety and effectiveness of their own device built on a foundation model.

B. Considerations for Agentic AI Systems

Agentic AI systems are GenAI-enabled systems that autonomously plan and execute multi-step tasks, use external tools, or take actions across a sequence of steps. They are increasingly being deployed in healthcare contexts for care coordination, clinical documentation, patient outreach, and clinical workflow support, some or all of which may not be functions that are the focus of FDA's device regulatory oversight. However,

[23] For more information, see FDA webpage Device Master Files.

[24] These may include, for example: the model's architecture, training data provenance, and intended supported use cases; characterized behaviors, known limitations, and failure modes relevant to healthcare contexts; evaluation results on healthcare-relevant benchmarks, including performance across clinically meaningful subgroups; safety-relevant behavioral constraints and guardrails built into the model; update notification commitments; and audit log availability.

there could be agentic AI systems whose functions meet the definition of device, for example, if an agentic AI system's action sequences result in control of another medical device.[25] Agentic architectures may raise additional considerations beyond those applicable to other GenAI-enabled devices that may warrant different approaches to evaluation. FDA seeks stakeholder comment on potential additional considerations that may be relevant for agentic GenAI-enabled devices.

C. Discussion questions

25. Would voluntary Foundation Model MAFs be practical for foundation model developers to provide and to keep current, and what content might they include? Given that participation would be voluntary and that model developers may have limited incentive to disclose safety-relevant information, what would make such a program sufficiently useful for premarket review? Are there alternative mechanisms CDRH should consider for obtaining information about underlying foundation models?
26. Are there additional considerations that inform the premarket and postmarket evaluation of agentic GenAI-enabled devices, beyond those applicable to non-agentic GenAI-enabled devices? How should the elevated risk associated with autonomous multi-step action, tool use, and reduced opportunity for human review be reflected in acceptance criteria and oversight?

VIII. Conclusion

The novel characteristics and behaviors of GenAI-enabled devices, including their capacity to produce variable outputs, evolve over time, and operate with varying degrees of autonomy, present potential challenges that existing regulatory frameworks were not designed to address. CDRH is committed to working collaboratively with sponsors, clinicians, patients, healthcare institutions, and other stakeholders to develop efficient, scientifically sound, and least burdensome approaches to the premarket evaluation and postmarket monitoring of GenAI-enabled devices across the TPLC. CDRH encourages stakeholders to respond to the discussion questions throughout this paper and to submit comments that may inform the development of further policy in this area.

[25] See Example #1 on Page 11 of FDA guidance Policy for Device Software Functions and Mobile Medical Applications.

Appendix A. Detailed Description of Potential Device Benchmarking Elements

This appendix describes in greater detail the potential benchmarking elements described in Section V.B. The descriptions below are illustrative of the kinds of behaviors each element may probe and the kinds of testing that may be relevant under a competency-based approach for premarket evaluation of GenAI-enabled devices described in this discussion paper. They are offered for stakeholder feedback and are not intended to prescribe any particular method.

Safety

These elements would evaluate whether the device avoids causing harm.

S.1. Safety-critical recognition and escalation. Whether the device recognizes safety-critical clinical states and initiates escalation that is timely, clear, actionable, and appropriately empathetic. Both under-escalation (failing to escalate when clinically necessary) and over-escalation (escalating when not clinically indicated) are relevant failures. Relevant testing may include multi-turn simulated encounters that assess time to escalation in evolving presentations, the quality of escalation language and handoff information, and resistance to over-reassurance when users minimize symptoms or resist escalation.

S.2. Scope maintenance and boundary adherence. Whether the device recognizes when a request exceeds its intended use or clinical scope and responds with appropriate refusal and clinically useful redirection. Both under-refusal (producing outputs beyond the device's intended use, such as diagnosing or prescribing when not designed to do so) and over-refusal (declining requests that fall within the device's intended use) are relevant failures. Relevant testing may include adversarial prompting, prompt injection, emotional-manipulation scenarios, and multi-turn conversations in which individual turns appear in-scope but the cumulative interaction drifts out of scope.

S.3. Calibration, uncertainty communication, and clinical deferral. Whether the device accurately represents the confidence and limits of its outputs, expressing appropriate uncertainty when evidence is contested, evolving, or beyond its reliable knowledge. Presenting uncertain, outdated, or contested information with false confidence is treated as a safety failure. This element also considers whether the device appropriately handles situations that are too complex or uncertain, including deferral to a human clinician, without over-deferring on questions within its knowledge.

Clinical Proficiency

These elements would evaluate whether the device possesses accurate clinical knowledge, reasons soundly through ambiguous information, interprets numerical data correctly, and communicates outputs clearly and appropriately to its intended users.

E.1. Clinical knowledge and task fidelity. Whether the device possesses accurate, current, and applicable clinical knowledge for its intended scope, including current guidelines, contraindications, drug interactions, and population-specific considerations, including, for example, those applicable to pregnant patients, pediatric and geriatric patients, and patients with renal or hepatic impairment. This element also considers whether the device is aware of its own knowledge limitations and acknowledges uncertainty for evidence not reflected in its training.

E.2. Information gathering and clinical analysis. Whether the device can gather, integrate, and act on incomplete, evolving, or ambiguous clinical information within appropriate clinical parameters. Relevant testing may include differential diagnosis generation assessed for completeness and appropriate prioritization of serious diagnoses, multi-turn simulated encounters scored for the quality and sequencing of follow-up questions, and scenarios designed to detect premature diagnostic closure in which initially benign information masks a more serious underlying condition. This element also considers whether the device defers appropriately when information is insufficient, without over-deferring when adequate information is available.

E.3. Quantitative and measurement analysis. Whether the device correctly interprets and acts on numerical and measurement information, including under the noisy or ambiguous input conditions common in real-world interactions. This includes clinically relevant calculations such as, for example, weight-based and renal-function-based dosing and unit conversions, interpretation of laboratory and physiologic values against appropriate reference ranges, and recognition of clinically implausible values such as transcription errors or unit confusion, including when measurements are reported conversationally or with missing units.

E.4. Communication quality and user comprehension. Whether the device communicates in a manner appropriate to its intended user, clinical context, and risk level, including empathy, linguistic construction, health literacy, and therapeutic appropriateness. This element considers whether intended users can actually understand the device's outputs, their limitations, and when escalation is needed across varied health literacy levels; the avoidance of coercive, emotionally manipulative, or otherwise inappropriate or abusive language; and the risk of automation bias, in which a user accepts an output without appropriate scrutiny because of the device's fluency or perceived authority.

Generalizability

These elements would evaluate whether safety and clinical proficiency properties hold consistently across foreseeable variation in inputs, users, and runtime conditions.

R.1. Robustness, reliability, and reproducibility. Whether the safety and clinical proficiency properties hold consistently across foreseeable variation in inputs, runtime conditions, and conversational contexts. Variation in non-safety-critical phrasing may be acceptable, but variation in safety-critical behaviors such as escalation, refusal, or diagnostic conclusions is treated as a failure. Relevant considerations include

reproducibility across repeated runs, consistency across semantically equivalent paraphrases reflecting differences in user health literacy and emotional framing, sensitivity to the order in which information is presented, behavior under missing or contradictory inputs, degradation across long conversations, and resistance to adversarial inputs targeting safety-critical behaviors.

R.2. Subgroup performance. Whether the safety and clinical proficiency properties hold consistently across clinically relevant populations, including populations that may be underrepresented in training data or face structural barriers to healthcare access. This element considers whether performance—including safety-critical behaviors such as escalation and refusal—is consistent across demographic subgroups, and whether it holds across non-standard dialects, accents, colloquialisms, and lower general literacy and health literacy levels.

Agentic AI Capabilities

This element would apply to agentic GenAI-enabled devices only.

A.1. Agentic AI competencies. Whether an agentic GenAI-enabled device appropriately plans, sequences, and executes multi-step tasks while recognizing when a planned action sequence would exceed its intended use or safety envelope. This element also considers accurate tool use and recognition of erroneous tool outputs, compliance with human-oversight checkpoints before irreversible or high-consequence actions, graceful handling of tool failures, and resistance to prompt injection through user inputs, retrieved content, and tool outputs.

Appendix B. Consolidated Discussion Questions

Discussion questions from the body of the paper, grouped by section:

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

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

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

7. Is the competency-based approach described above, i.e., device benchmarking followed by clinical confirmation, a useful and appropriate framework for evaluating GenAI-enabled devices?
8. How might the two-axis risk framework described in Section IV be considered within the competency-based approach described above to help determine the level of evidence needed for a premarket submission?
9. Would a benchmarking structure such as the one described above be likely to provide adequate evidence of clinical knowledge, analytic capabilities, safety behavior, communication, and generalizability to support a reasonable assurance of safety and effectiveness? Are there elements that are missing, redundant, or inappropriately categorized? Are there externally developed standards that could be leveraged?
10. CDRH recognizes that publicly available benchmarking assets may be subject to data contamination, saturation, and limited real-world representativeness. How should a sponsor establish that performance on a given benchmark predicts safe and effective real-world behavior for the device's intended use? What evidence should support the construct validity of a benchmark used to gate device evaluation, and what role should sponsor-developed benchmarks play given potential concerns around independence and optimization to the test?
11. CDRH is considering that clinical confirmation for a GenAI-enabled device might not require a prospective clinical study in every case, and has described above a range of approaches of increasing rigor and patient exposure. How might a sponsor select and justify a confirmation approach tailored to a device's intended use and proportionate to the device's risk profile? Are there device types or risk profiles for which one or more of these approaches would be insufficient or inappropriate? How might the anticipated distribution of real-world inputs be taken into consideration? Are there other methods of clinical confirmation that might help inform the evaluation of GenAI-enabled devices?
12. How should sponsors achieve statistically meaningful performance measurement for GenAI-enabled devices? Where synthetically generated inputs supplement real patient data, how should sponsors account for differences between the synthetic and real-world distributions when estimating performance, and under what conditions, if any, is it appropriate to combine benchmarking evidence and clinical confirmation evidence to support a single performance estimate?
13. For which clinical domains, device functions, or subpopulations is synthetic data particularly well-suited, or particularly inadequate, as a supplement to real-world evidence? What safeguards would mitigate the risk that synthetic data generated by models of the same class as the device under evaluation reproduces the very performance gaps the evaluation is intended to detect, particularly for underrepresented subgroups?

14. For open-ended device outputs, how should performance comparators and acceptance criteria be selected? When a panel of qualified clinicians serves as the comparator, how should the applicable standard (for example, the standard of care versus the performance of a median clinician in practice) be defined and justified? Should generalist or specialist physicians be used as a performance standard? When should human-AI team performance, rather than the device operating alone, serve as the basis for evaluation?
15. Are there ways in which performance might be assessed relative to the care, technology, or course of action likely to occur in the absence of the device, rather than to the comparators described in this section? What approaches might be used to identify and justify a comparator such as unaided clinical judgment, delayed specialist review, or no intervention?
16. Should independent third parties be involved in some or all aspects of a competency-based approach, including device benchmarking and clinical confirmation? If so, in what ways might qualified, independent third-party participation contribute to this approach, and what qualifications and independence criteria should apply? Are there aspects of the assessment for which third-party involvement would be impractical or inadvisable? What safeguards or program design features would be critical to prevent third-party participation from limiting competition or preventing innovation?
17. CDRH is exploring whether a competency-based approach could be applied to devices that incorporate a variety of underlying model architectures, such as multimodal vision-language models and generative or predictive world models. Are there aspects of the described approach that might be ineffective or inapplicable to these devices? What additional or distinct considerations might need to be addressed?

Section VI: Postmarket Monitoring for GenAI-Enabled Devices

18. CDRH is considering whether it may be appropriate to accept greater premarket uncertainty regarding a GenAI-enabled device's benefit-risk profile through greater reliance on postmarket monitoring. Under what conditions might such an approach be appropriate, and what characteristics of a monitoring program would need to be in place to justify reduced premarket evidence? Are there device types or risk profiles for which this approach would not be appropriate?
19. Please comment on the potential approaches to postmarket performance evaluation, including periodic re-benchmarking, sample-based clinician review, and performance degradation monitoring. What additional approaches should CDRH consider, and how should the cadence and triggering events for reassessment be determined?
20. Please comment on whether the proposed approaches to postmarket monitoring can be facilitated by machine-based supervisory agents. What considerations, including the evaluation and reliability of the supervisory agent itself, should CDRH take into account for such an approach?
21. What roles might clinicians, healthcare institutions, professional societies, standards-setting bodies, and other stakeholders appropriately play in

postmarket monitoring, and how can these roles be structured without diffusing manufacturer accountability?

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?
23. How might PCCP concepts or other change-control approaches be adapted for GenAI-enabled devices when the nature or scope of future modifications cannot be fully prespecified?
24. For devices built on third-party foundation models, changes to the underlying model may be initiated by the third-party model developer rather than the device manufacturer. How can a manufacturer detect, evaluate, and respond to such changes in a timely manner? What mechanisms—for example, contractual, technical, or through a PCCP—could provide reasonable assurance that third-party developer-initiated changes do not compromise the safety or effectiveness of the device?

Section VII: Other Topics

25. Would voluntary Foundation Model MAFs be practical for foundation model developers to provide and to keep current, and what content might they include? Given that participation would be voluntary and that model developers may have limited incentive to disclose safety-relevant information, what would make such a program sufficiently useful for premarket review? Are there alternative mechanisms CDRH should consider for obtaining information about underlying foundation models?
26. Are there additional considerations that inform the premarket and postmarket evaluation of agentic GenAI-enabled devices, beyond those applicable to non-agentic GenAI-enabled devices? How should the elevated risk associated with autonomous multi-step action, tool use, and reduced opportunity for human review be reflected in acceptance criteria and oversight?