

October 19, 2026

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

Re: (Docket ID: FDA-2026-N-7874) Considerations for the Regulation of Generative AI-Enabled Medical Devices: Discussion Paper and Request for Feedback; Comments of the American College of Radiology

The American College of Radiology (ACR) – a professional association representing more than 40,000 physicians practicing diagnostic radiology, interventional radiology, radiation oncology, and nuclear medicine, as well as medical physicists – appreciates the opportunity to comment on the discussion paper published August 18, 2026, by the Food and Drug Administration (FDA) titled *Considerations for the Regulation of Generative AI (GenAI)-Enabled Medical Devices*. ACR previously presented on this subtopic in-person to the FDA Digital Health Advisory Committee during its public meeting on Nov. 20-21, 2024, and filed written comments on Jan. 21, 2025, during the subsequent comment period corresponding with the same meeting. ACR supports FDA's continued efforts to gather stakeholder perspectives on this topic before proposing any new policy.

General Comments

Radiology is a leading specialty domain for GenAI-enabled medical device innovation, and image-informed draft reporting is among the most active areas of development and early clinical evaluation. A central challenge for FDA oversight of GenAI-enabled devices in radiology and other specialties is that device performance often depends not only on the software itself, but also on the clinicians, workflows, data sources, governance structures, and monitoring systems with which it operates.

The ACR's complementary approach to post-deployment AI safety—including continuous post-market [performance monitoring](#), future practice accreditation for healthcare facilities, and the [responsible use](#) of radiology AI in practice—has the ability to pick up on this composite system performance, particularly as these safeguards become increasingly connected to longitudinal outcomes data. We welcome continued partnership between FDA and the medical community to take on these challenges.

Consistent with this view, we urge FDA to incorporate the following principles into any future oversight framework for GenAI-enabled medical devices:

- Regulate what the system does and the risk it creates. FDA should regulate the complete clinical function and its associated risk. The same underlying foundation model may support low-risk documentation or high-risk autonomous diagnosis. As a result, regulatory requirements should reflect intended use, potential patient harm, workflow, and degree of autonomy.

- Explicitly credit qualified end-user oversight, but only if demonstrated. A GenAI output reviewed and affirmatively accepted by a qualified end-user before the output influences patient care has a fundamentally different risk profile from an output that independently makes, communicates, or acts upon a clinical determination. FDA should therefore recognize qualified-user review as an important risk control, but only while accounting for error detectability, automation bias, and the user's ability to independently assess the output. This oversight should earn credit only when sponsors demonstrate its effectiveness in the intended workflow (e.g., error detection and correction under realistic workload, time pressure, and interface conditions, with residual errors reported after review).
- Allow technology to evolve. FDA should move toward validated performance envelopes, predetermined change control plans, and lifecycle oversight where these mechanisms are appropriate. Changes to models, prompts, retrieval, orchestration, or guardrails that remain within validated performance parameters should not automatically require a new submission provided safety and effectiveness remain the same.
- Counterbalance premarket flexibility with real-world performance monitoring. Greater flexibility for continuous improvement should be counterbalanced by robust postmarket monitoring for clinically meaningful failures, performance drift, subgroup effects, and emerging risks. For GenAI, postmarket monitoring should be viewed not simply as an additional requirement but as a core component of the regulatory framework that enables safe continuous improvement.
- Measure the failures that matter clinically. Evaluation should focus on clinically consequential performance rather than generic hallucination rates, nondeterminism, or textual similarity. ACR and other medical societies can help define meaningful categories such as significant false assertions, omissions, laterality/localization errors, incorrect comparisons, unsupported recommendations, contradictions, and communication failures. GenAI requirements should also scale with autonomy.
- Clarify relevance to already-authorized medical devices. In any future regulatory framework specific to GenAI, FDA should clarify applicability to currently authorized medical device functions, such as those enabling generative image reconstruction, denoising, and synthesis.
- Add worked imaging examples. Given the level of innovation in GenAI-enabled medical devices within the radiology domain, any future regulatory framework, guidance, and other FDA resources should include imaging examples.
- Seek imaging community feedback via dedicated activities. The discussion paper is broadly applicable across all GenAI-enabled medical device issues. To better address imaging, FDA should convene a radiology-focused public workshop, Digital Health Advisory Committee meeting, and/or publish a targeted solicitation for written comments on imaging-specific considerations prior to establishing a future regulatory framework, developing draft guidance, or other policy developments.

Specific Questions

Q1. 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.

A1. The described two-axis risk framework, which evaluates device activity and the consequences of relying on an incorrect output, may oversimplify or overlook important contextual considerations that influence risk. Risk in most clinical applications is not a function of the AI system alone, but of the broader human-AI/socio-technical system in which it is deployed and used. Exceptions may include highly autonomous AI functions or narrowly defined functions such as workflow triage, image transformation, or text transformation that serve as individual components within a larger clinical system.

We recommend adding contextual modifiers that refine risk assessments within the framework rather than adding dedicated axes that may complicate the framework without adding clarity. These considerations should include:

- Severity of harm from wrong outputs: Delineation of different levels of potential harm. FDA describes severity with examples, but does not define the Limited, Moderate, and Severe levels of the consequences axis.
- Expected frequency of wrong outputs: The frequency of erroneous outputs produced by the device across the expected input mix and conditions of use.
- Detectability: Whether, and at what point, the intended user detects an erroneous output within the real-world workflow. Different users may act differently when they detect an error, which may impact subsequent actions and harm. Poor detectability increases the likelihood that an erroneous output will result in harm and may increase the number of affected patients before the user or system identifies the issue.
- Safeguards/deployment exposure: Separate consideration of device safeguards and what deploying organizations must do to ensure safety. Rather than merely requiring description of safeguards, require evidence that safeguards are effective under expected conditions of use.
- Reversibility: Whether there is a practical and meaningful opportunity for human intervention to undo the action between detection of an erroneous output and its effect on patient care.
- Traceability/source attribution: The ability to identify information underlying an output and reconstruct decisions with confidence that citations are accurately and correctly attributed. An erroneously generated or misattributed source is itself a failure mode that can exacerbate safety issues via persuasiveness.

Finally, any future FDA guidance or other resource that establishes this framework as agency policy should include axis-placement criteria and examples, including examples of low-autonomy functions with severe downstream consequences. Risk assessment of GenAI will require greater nuance than the discussion paper can illustrate via an abstract, generalized framework, and therefore providing more examples would enhance clarity.

Q2. 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?

A2. The characteristics of an output described in question 2 are appropriate, but operationally challenging. ACR supports FDA’s notion in the discussion paper that disclaimer content in a patient-targeted output does not neutralize the core directive in that output.

Regarding context as an output characteristic, the key question in imaging is whether the reader can independently assess the output and identify consequential errors. That assessment typically lowers risk without establishing “low risk.” Harm severity if an erroneous output goes undetected, reader time pressures, directly or indirectly triggered or suggested action(s), and presentation still matter, however. The FDA should require the sponsor to demonstrate the detectability of errors, including whether readers catch errors under real-world workloads and within vendor-provided interfaces. The sponsor should also produce those measurements.

Personalization of the output can increase risk. For patient-users, personalization can make the output seem more directive and detailed, which raises risk level. For clinician-users, a personalized output, such as matching a radiologist’s phrasing, can lower editing burden, but it makes the AI output read like the radiologists’ own, which can lower scrutiny by the reporting physician. FDA should evaluate detection performance in the personalized setting, not exclusively in the default setting.

Finally, FDA should clarify whether directiveness is a characteristic of the device function or of a particular output. This distinction is especially important for conversational systems, whose outputs may range from purely informational to highly directive depending on user prompts.

Q3. 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?

A3. FDA should consider the functions and safeguards in place for patient-delivered outputs, rather than defaulting to higher risk determinations merely because the patient is an intended user. Patient users have widely varying health literacy levels, and this changes the risk assessment paradigm of many functions due to different error detection capabilities, output reliance, downstream action, and time-to-treatment. In addition, the potential risk is impacted by the intended use of an AI product. For example, a CT prep instruction function has different risk considerations than a symptom-triage function in practical terms. As both factors impact the probability of harm and its severity in different ways, FDA should separate the two rather than putting both on the consequences axis.

Safeguards should improve accuracy, uncertainty management, and user comprehension, rather than rely primarily on users observing and adhering to disclaimers. Examples include grounding responses in current and authoritative information sources, transparency regarding retrieved content and its limitations, validated approaches for communicating uncertainty with appropriate escalation thresholds, and demonstrated comprehension across the intended user population, including patients with varying levels of health literacy. For higher-risk clinical decisions, systems should explicitly avoid directing patients to initiate, discontinue, or modify treatment without appropriate healthcare professional oversight.

Q4. 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?

A4. The answer depends primarily on the specific function's intended use, potential consequences of error, available safeguards, and whether the intended user can independently recognize and appropriately respond to consequential errors.

The 2026 [*ACR–SIIM Practice Parameter for Imaging Artificial Intelligence*](#) defines a “qualified end-user” in this context as “a licensed physician with the necessary qualifications and training to independently provide the equivalent medical service without the aid of the AI solution, and who possesses qualifications and training in the use of the AI solution, including the ability to assess the validity of its output.” FDA should recognize qualified end-user review as an important risk control when the user evaluates and affirms the output before it influences patient care decisions or actions. However, the effectiveness of this safeguard depends on the user's expertise, the detectability of errors, the potential for automation bias, and whether specialist consultation or referral is readily available in a timely fashion and effective within the intended workflow.

Additional safeguards might include output-uncertainty communication, recommendations for specialist review or escalation in uncertain or high-risk situations, and system-level performance monitoring.

Q5. 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?

A5. The paper discusses drift symptoms across three benchmark items but leaves these unconnected. For example, S.1 discusses “resistance to over-reassurance when users minimize symptoms or resist escalation.” S.2 asks about “multi-turn conversations in which individual turns appear in-scope, but the cumulative interaction drifts out of scope.” R.1 lists degradation across long conversations.

Sponsors should define intended use across the conversational trajectories the device supports, including the transition from informative outputs to directive or suggestive outputs.

Informational/non-directive claims need evidence that the device holds that boundary. Where safety depends on specialist escalation or user scope limitations, multi-turn adversarial testing under sustained user pushback and the presence of sponsors to measure whether the device withdraws or softens a warranted escalation with prespecified acceptance criteria should be required.

Q6. 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?

A6. ACR recommends evaluating under-escalation and over-escalation as distinct performance dimensions, with separate validation criteria, risk assessments, and mitigation strategies for each. For patient-facing or patient-directing AI, sponsors should clearly communicate risks associated with both under-escalation and over-escalation.

The acceptable balance between under-escalation and over-escalation is highly context dependent. In medical imaging, for example, over-calling an actionable incidental finding may lead to additional imaging, radiation exposure, healthcare costs, and patient anxiety. Once embedded in a clinical report, an unnecessary but documented recommendation may also be difficult for clinicians to reverse. Conversely, under-calling a finding may delay diagnosis or treatment. As demonstrated by the ACR Incidental Findings Committee's work, determining appropriate trade-offs requires consideration of the specific clinical context and evidence base.

FDA should therefore require sponsors to report under-escalation and over-escalation separately, with predefined acceptance criteria for each. Combining the two into a single aggregate escalation metric could create incentives to offset under-escalation with over-escalation conservatism.

Q7. 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?

A7. ACR supports a framework in which the sponsor follows benchmark testing with clinically grounded confirmation of performance for the device's intended use. Benchmark evaluation can establish a minimum performance threshold, but benchmark results alone do not demonstrate that a device will perform safely and effectively within diverse real-world clinical settings, populations, or workflows. The sponsor should tie clinical confirmation to the device's intended use, target population, and conditions of use.

FDA should avoid relying on aggregate benchmark scores alone. Benchmark performance may be sensitive to prompt wording, question ordering, repeated runs, and other testing conditions, and high benchmark performance does not necessarily predict safe and effective performance for the intended use, population, or workflow. Sponsors should therefore provide evidence that benchmark results are predictive of real-world performance for the intended use and supplement summary metrics with sensitivity testing, failure analysis, and evaluation of clinically consequential errors.

Additionally, if FDA intends to use ongoing benchmarking as part of a lifecycle oversight approach, the agency should provide clear expectations regarding performance monitoring, re-evaluation triggers and thresholds, remediation requirements, and appropriate regulatory actions when the regulator, sponsor, or other stakeholder identifies performance degradation.

Finally, FDA should exercise caution when drawing analogies between GenAI evaluation and the education, training, or credentialing of healthcare professionals. Competency is a characteristic of individuals, whereas software medical devices are evaluated based on demonstrated performance for defined tasks and intended uses. While such analogies may be useful for illustration, regulatory requirements should remain focused on evidence supporting the device's safety and effectiveness for its intended use.

Q8. 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?

A8. ACR agrees that evidence requirements should vary based on the intended use, level of device activity, and potential consequences of reliance on incorrect outputs. If FDA adopts the proposed framework, the agency should first establish clear placement criteria and provide illustrative examples before proposing corresponding evidence requirements. Evidence expectations should then use the contextual risk modifiers described in our response to question 1.

For imaging applications, ACR believes FDA should require clinical confirmation whenever the intended user cannot independently assess the accuracy of the output, regardless of where the function falls on the activity axis of the framework. This distinction helps ensure that the agency does not treat functions, such as generative image reconstruction, that indirectly inform diagnoses as low-risk informational tools solely because they do not directly offer diagnostic recommendations. Instead, the level of evidence should reflect the user's ability to detect consequential errors and the potential downstream impact on clinical decision-making.

Q9. 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?

A9. FDA should evaluate failures that matter clinically rather than generic hallucination rates, nondeterminism, or textual similarity. Medical societies like ACR can help define meaningful categories, and any such requirements should also scale with autonomy. To that end, there are two gaps that are missing or inadequately addressed with respect to imaging GenAI specifically—an imaging failure taxonomy and a scoring architecture.

With respect to compiling a failure taxonomy, this must have testability, operational definitions, severity classifications, reference evidence per category, and thresholds. FDA should partner with medical societies, such as ACR, to build and maintain that taxonomy for relevant subdomains. For imaging, clinically meaningful failures can include:

- laterality errors,
- localization errors: correct finding, wrong site, or wrong level,
- fabricated or incorrect comparison to a prior, including comparison to a study that does not exist,
- measurement hallucination, or a correct measurement attached to the wrong lesion,
- omission of a clinically significant incidental finding,
- incorrect study, protocol, or technique attribution,
- report-image discordance,
- failure to communicate critical or actionable results,
- erroneous or absent follow-up recommendations, including recommendations inconsistent with published guidelines,
- anatomic mislabeling, and,
- for reconstruction and synthesis: structures introduced that are not in the acquired data, and true findings lost or attenuated.

Regarding scoring architecture, the discussion paper does not clearly indicate whether FDA might intend for performance assessment to be compensatory or non-compensatory. As written, strong performance in one domain could offset poor performance on clinically important safety measures within a composite score. ACR recommends that clinically important safety measures carry prespecified minimum performance thresholds that cannot be offset by strong device performance in unrelated evaluation domains. This does not mean that every observed safety error should result in automatic failure; rather, FDA should define appropriate endpoints, performance thresholds, and uncertainty bounds. The goal should be to avoid aggregate scores that mask unacceptable safety-related performance.

FDA should also clarify the regulatory implications of failing to meet a safety threshold, including whether the agency would require additional evidence, risk controls, labeling restrictions, narrower intended uses, or other actions. The relevant regulatory question is whether the device's demonstrated performance supports a favorable benefit-risk profile for the proposed intended use.

With respect to external standards and clinical guidelines, FDA should encourage the use of independently developed, specialty-specific standards where they exist. For medical imaging, the ACR Appropriateness Criteria® are an example of the guideline-based frameworks described in the discussion paper. Developed through evidence review, multidisciplinary expert input, and structured consensus processes, such resources can provide clinically grounded reference standards for evaluation rubrics, performance measures, and assessment of GenAI outputs. Existing experience using the ACR Appropriateness Criteria as a reference standard for evaluating LLM-generated imaging recommendations also demonstrates how established clinical guidelines can support more clinically meaningful assessment than generic benchmark metrics alone.

Q10. 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?

A10. ACR agrees that benchmark datasets should be representative of the intended use, patient population, input data, and real-world deployment environment. Clinical confirmation demonstrating predictive value for real-world performance should corroborate benchmark performance.

For imaging applications, FDA should encourage the use of both clinically representative patient data and appropriately validated data phantoms. Patient data are necessary to evaluate performance under real-world conditions, while phantoms can provide controlled reference standards for specific technical characteristics for which assessment using clinical cases alone is inadequate.

Assessments may use sponsor-developed benchmarks for evaluating specialized use cases or addressing gaps not covered by publicly available resources. Nonetheless, these benchmarks should not serve as the sole basis for demonstrating safety or effectiveness, particularly for higher-risk functions. FDA should consider requiring independent evaluation datasets, independent clinical review, or other external validation measures when sponsors use benchmark results to support higher-risk claims. At a minimum, sponsors should provide transparent documentation regarding benchmark construction, evaluation methodologies, adjudication processes, acceptance criteria, known limitations, and measures taken to maintain independence between development and evaluation activities.

FDA should also address potential overlap between training and evaluation data. While the extent of benchmark contamination in medical AI evaluation may not always be known, sponsors should disclose known overlaps, characterize remaining uncertainty, and describe measures taken to reduce bias in performance estimates. Additional independent, temporally separated, or real-world clinical datasets may be needed if meaningful overlap cannot be excluded. Assessors should also periodically refresh or update benchmark datasets to preserve their long-term validity as evaluation tools.

Q11. 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?

A11. The rigor of clinical confirmation should increase with the degree of autonomy, reliance on the output, and potential patient impact. Functions that are autonomous, patient-facing, or otherwise intended for use without effective qualified end-user oversight warrant more rigorous clinical confirmation than functions whose outputs are qualified end-user-vetted before influencing care.

FDA should also clarify that clinical realism, evaluation rigor, and exposure to device-influenced care are distinct concepts that do not necessarily increase together. For example, a shadow deployment may involve no exposure to device-influenced care while still evaluating performance

on real patients in a real clinical workflow. Separating these dimensions would improve consistency in selecting appropriate clinical confirmation approaches.

Sponsors should define the anticipated deployment environment and demonstrate adequate coverage of the expected input distribution, including patient demographics, care settings, equipment manufacturers and models, acquisition protocols, and other variables material to the intended use.

Finally, FDA should require clinical confirmation using real patient data whenever the intended user cannot independently assess the accuracy of the output. This consideration may be particularly important for functions involving generative reconstruction, synthesis, or other outputs where clinically consequential errors may not be readily detectable.

Q12. 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?

A12. Sponsors should anchor clinical performance claims in representative real-world patient data. Synthetic data, digital phantoms, and other simulated inputs may support development, testing, and certain validation activities, but their evidentiary value depends on the claim being supported, the availability of known reference truth, and their representativeness of the intended use environment. Accordingly, FDA should require that the sponsor separately evaluate performance on real-world and synthetic data.

Regardless of data source, FDA should require sponsors to: (1) prespecify clinically relevant endpoints, acceptance criteria, sample-size justification, confidence bounds, and statistical analysis methods; (2) clearly define the unit of analysis and account for dependence across patients, sites, readers, conversations, or repeated model runs; (3) prespecify relevant subgroups and evaluate clinically consequential subgroup claims on representative data; and (4) document the deployed system configuration, including model version, prompts, retrieval mechanisms, guardrails, and other components that may influence performance.

FDA should also address performance variability. For GenAI systems, identical inputs may produce different outputs across repeated runs. Sponsors should therefore assess variability in clinically meaningful conclusions and safety-related outcomes using prespecified acceptance criteria and uncertainty bounds. Variability reporting should inform the safety and effectiveness assessment but should not substitute for meeting performance requirements.

Finally, benchmarking and clinical confirmation serve different evidentiary purposes, and sponsors report these separately. Combining results may be appropriate only when sponsors justify comparability, weighting, and statistical assumptions, and demonstrate that the resulting estimate remains representative of the intended use population and deployment environment.

Q13. 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?

A13. Representative real-world patient data should remain the primary basis for demonstrating clinical performance. Synthetic data appropriateness depends on the claim the sponsor intends for that data to support. FDA should avoid treating all synthetic data as a single category.

Different uses of synthetic data warrant different evidentiary standards:

- Physics-based simulations and digital phantoms may provide controlled reference truth but require validation of their assumptions and realism.
- Rare-event, adversarial, and stress-testing applications may reveal important failure modes but do not establish real-world performance or prevalence.
- Training augmentation should be evaluated based on performance improvements demonstrated on independent real-world datasets. Such evidence supports the value of the augmentation but does not establish the absence of unintended effects.
- Clinical performance assessment using synthetic data should require a justified relationship to the intended-use population, with synthetic and real-world results reported separately.

FDA's concern should extend beyond reproducing known performance gaps. Synthetic data generation may introduce, amplify, suppress, or otherwise alter clinically relevant attributes that were not the intended target of manipulation. Accordingly, sponsors should evaluate potential unintended changes relevant to the proposed use, characterize the limits of those analyses, and disclose remaining uncertainty.

Sponsors should also disclose the synthetic data generation process, including the generator, available information regarding training-data provenance, the relationship between the generator and the evaluated device, and any material uncertainties they cannot resolve. If sponsors generate synthetic data recursively or to augment future training datasets, they should assess and document potential impacts on diversity, representativeness, and performance.

Finally, clinically consequential subgroup claims should be confirmed with independent real-world data. Sponsors should also demonstrate that their evaluation approach can detect relevant device failures despite potential dependencies arising from shared models, datasets, assumptions, or evaluation tools.

Q14. 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?

A14. FDA should distinguish among three separate questions that may require different evidence: (1) whether an output is clinically acceptable, (2) how performance compares to existing clinical practice, and (3) whether the device has met minimum safety and effectiveness requirements. These considerations can coexist, but one should not substitute for another.

Clinical acceptability should be evaluated against either an objective reference standard where one exists (e.g., pathology, adjudicated outcomes, or follow-up imaging) or through independent expert adjudication using predefined criteria. For open-ended outputs, multiple responses may be clinically acceptable, and qualified expert panels may be better suited to defining acceptable ranges of performance than a single "correct" answer.

Comparator selection should reflect the intended use, deployment setting, and clinician who would otherwise perform the task without the device. For imaging interpretation, the appropriate comparator would typically be the radiologist who would otherwise perform that function. FDA should clarify the distinction between "standard of care" and "median clinician performance," as these concepts are not equivalent and may lead to different conclusions regarding benefit and risk. Sponsors should justify the selected comparator based on the clinicians and workflows the device is intended to support.

Finally, device-only performance and human-AI team performance are separate considerations whenever performance depends materially on user interaction, oversight, or workflow integration. Standalone device accuracy does not necessarily predict performance within a clinical workflow, particularly for GenAI functions where automation bias, plausible omissions, or editing burden may affect real-world use. The extent of workflow evaluation and human oversight should be proportionate to the device's intended use, autonomy, and risk profile.

Q15. 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?

A15. The value of a GenAI-enabled device should be assessed relative to the care that would realistically occur in the absence of the device. The appropriate comparator may vary across intended uses and deployment settings. In some environments, the relevant comparator may be prompt specialist interpretation; in others, it may be a delay of specialist review, preliminary review by another clinician, or no timely interpretation. As a result, the same device performance may produce different benefit-risk profiles across different practice settings.

FDA should therefore require sponsors to justify the selected counterfactual using evidence from the intended deployment environment, including existing workflows, practice patterns, resource availability, and relevant clinical data. The sponsor should include in intended use/labeling all claims based on delayed, limited, or absent specialist access.

At the same time, a favorable counterfactual should not substitute for independent demonstrations of accuracy, safety, and effectiveness. Assessments of clinical acceptability, comparisons to existing care, and minimum safety assurances address different regulatory questions, and each should be evaluated separately.

Finally, because the validity of a counterfactual depends on the deployment environment, FDA should consider postmarket monitoring when authorization relies on assumptions regarding site capabilities, specialist availability, or clinical workflow. Continued use outside the settings supporting the original benefit-risk assessment may warrant additional evaluation.

Q16. 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?

A16. ACR supports the use of qualified, independent third parties for benchmarking, clinical confirmation, and postmarket assessment. For purposes of this comment, we define a qualified third-party as “a trusted, unconflicted entity that is independent of both the manufacturer and the regulator.” At a minimum, FDA should require third-party assessors to have the following characteristics of independence:

- disclosure of relevant financial relationships at both the organizational and individual reviewer levels.
- individual recusal requirements with an FDA-defined look-back period.
- transparent funding arrangements insulated from assessment outcomes.
- organizational separation between assessment activities and related commercial services provided to the same sponsors.
- prespecified evaluation protocols, reporting of results regardless of outcome, and no sponsor right to suppress findings.
- prohibition on manufacturers evaluating their own devices or those of competitors; and
- multiple qualified evaluators rather than a single gatekeeper.

FDA should require manufacturers to make device outputs and relevant performance information (e.g., AI outputs and related data) available in standardized, interoperable formats (e.g., DICOM Structured Reports or HL7 FHIR) that support timely independent evaluation with near-real-time release rather than batch releases. Without such access, meaningful and timely third-party assessment is not feasible.

ACR's [Assess-AI](#) initiative, the first nationwide imaging AI data service including a registry function, represents existing postmarket monitoring infrastructure that may help inform future independent assessment approaches. ACR would also support development of independently governed evaluation datasets and adjudication services, with validation, governance, funding, and operational capacity established before such resources were used by sponsors/FDA to support regulatory decisions. ACR's experience with Assess-AI, AI Central, Transparent-AI, and ARCH-AI recognition may provide useful perspectives in this area. However, any organization serving a third-party assessment role should be subject to the same independence, governance, transparency, and conflict-of-interest requirements.

Finally, FDA should distinguish between assessment, certification, accreditation, and FDA authorization of devices, as these activities serve different purposes. Certification-style examinations or sequestered testing assets may help address benchmark contamination concerns, but they do not by themselves establish safe and effective performance of a device in clinical practice. Performance on clinically relevant real-world tasks should remain the primary basis for regulatory decision-making.

Q17. 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?

A17. FDA should clarify how the framework applies to generative reconstruction, enhancement, and image synthesis functions. While these functions may not directly provide diagnostic conclusions, they generate or modify the images upon which subsequent clinical decisions depend. As a result, a function's apparent level of activity may not accurately reflect its risk.

For reconstruction and synthesis functions, FDA should require evaluation using datasets with known reference truth, including appropriately validated phantoms, inserted lesions, or similar methods. Assessment should specifically evaluate whether the function introduced, altered, obscured, or lost clinically relevant findings.

Additionally, FDA should not assume the reading radiologist will always catch a reconstruction error. Instead, device-specific evidence should be required before crediting visual review as a safeguard.

Q18. 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?

A18. In certain circumstances, greater premarket uncertainty may be acceptable when counterbalanced by robust postmarket monitoring. However, the appropriateness of this approach depends on the device's intended use, degree of autonomy, availability of qualified end-user oversight, and the potential consequences of error. This tradeoff is more defensible when a qualified end user reviews each output before it influences patient care and is less appropriate for patient-facing or autonomous functions.

Deferral is appropriate only when postmarket monitoring can reliably detect clinically meaningful failures, performance drift, subgroup effects, and emerging risks in time to prevent or limit patient harm. Sponsors should identify the signals they are monitoring, the underlying data sources, their expected detection timelines, and corrective action plans.

FDA should require prespecified performance thresholds, remediation timelines, notification procedures, and conditions for deployment restrictions or withdrawal. Sponsors should also establish workable mechanisms for obtaining the data necessary to support monitoring and demonstrate that adequate resources exist to sustain those activities.

Finally, greater reliance on postmarket evidence may be inappropriate where unresolved uncertainty involves the potential for serious, irreversible, or catastrophic harm. Even when such harms are theoretically detectable through monitoring, the inability to intervene before meaningful patient exposure occurs may warrant additional premarket evidence.

Q19. 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?

A19. ACR supports the use of re-benchmarking, sample review, and drift monitoring as complementary components of a lifecycle oversight framework. These approaches serve different purposes, and FDA should not consider these interchangeable. Re-benchmarking can identify performance regression against established evaluation criteria, sample review can identify unanticipated failure modes, and drift monitoring can identify changes in the population, clinical environment, or device behavior over time.

Effective monitoring requires access to real-world clinical data and workflows that are often not available to manufacturers alone. Therefore, partnerships with organizations that hold those data are necessary. ACR's Assess-AI provides existing infrastructure for this purpose. Participating sites submit de-identified AI outputs, report text, and study metadata through locally installed software. Concordance against report-derived surrogate labels is then computed centrally, and sites receive longitudinal performance trends and benchmark comparisons and can utilize local re-identification capability for case review. Assess-AI currently supports classification use cases such as intracranial hemorrhage, pulmonary embolism, and large-vessel occlusion. ACR intends to extend this architecture to GenAI outputs in the future.

Monitoring programs should capture all eligible cases, including processing failures, missing outputs, abstentions, raw outputs before clinician editing, and clinically relevant downstream outcomes. At the same time, data completeness and other limitations must be transparently reported. Sponsors should also conduct impact assessments for material changes to models, prompts, retrieval systems, orchestration components, guardrails, or other elements that may affect performance.

Finally, reassessment should occur on a risk-informed basis, triggered by predefined events, including safety signals, performance drift, significant model or workflow changes, or established time intervals. Assessors should periodically refresh or rotate evaluation datasets to reduce overfitting from repeated testing. Disclosures of important methodological limitations, including imperfect reference standards, incomplete data capture, and validation efforts that remain ongoing, should accompany the monitoring results.

Q20. 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?

A20. Machine-based supervisory agents will be necessary to evaluate and monitor GenAI-enabled devices at the scale of modern healthcare workflows, including medical imaging. However, FDA should treat supervisory systems as safety-critical components subject to independent validation and ongoing oversight. The acceptable degree of machine supervision should vary based on the device's intended use, degree of autonomy, and whether the device is patient-facing or provider-facing.

We caution that LLM-based evaluators may exhibit preferential scoring, including favoring outputs produced by similar models or by the evaluator itself. Agreement between machine supervisors and human reviewers does not necessarily establish safety and reliability. Moreover, the supervisor is an attack surface, and short adversarial phrases appended to a text can push judge LLMs to maximum scores regardless of content. Therefore, FDA should require the following:

- Disclose the supervisor's base model, lineage, and available training provenance, including what is unknown. Shared lineage warrants scrutiny, and different lineage does not establish independence. Sponsors should have to measure correlated failure of the supervisor.
- Evaluate for rare consequential failures and abstention behavior rather than overall agreement. A supervisor that never identifies uncertainty may not be functioning as an effective reviewer.
- Measure how often the supervisory system fails to identify consequential device errors and whether the addition of supervision meaningfully improves safety relative to the unsupervised device.
- Validate the adjudication method against explicit clinical acceptability criteria, including cases where every candidate output is unsafe. Test adversarial and out-of-distribution inputs before the sponsor uses the supervisory system to support safety or effectiveness claims.
- Apply the same drift monitoring as the device, and have humans audit a defined sample of the device's judgments on an ongoing basis.
- For panels, measure joint failures and the panel's added value over single evaluators. Neither the number of agents nor lineage diversity establishes an adequate control.

Q21. 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?

A21. Professional societies should define what gets measured and maintain the measurement tools. For example, in radiology, ACR can contribute resources with distinct roles, such as Assess-AI for post-deployment surveillance and future practice accreditation for deploying healthcare facilities. This strengthens rather than diffuses manufacturer accountability by ensuring someone other than the manufacturer is also monitoring performance. Section VI.B recognizes these complementary roles, but does not specify how FDA would assign measurement, reporting, and

corrective-action responsibilities without shifting the manufacturer's obligations onto clinicians and institutions. FDA should clarify the above consideration in any future relevant policy.

Additionally, the stakeholder ecosystem is more complicated than suggested in the discussion paper and question 21. In imaging, one organization may acquire the images, a second may employ the interpreting radiologist, and a third may operate the AI. For each deployment, stakeholders must assign named responsibility for investigating alerts, stopping use, notifying affected sites, reviewing cases already affected, and maintaining care during outages. Such stakeholders should settle data-sharing and escalation considerations before going live. Manufacturer accountability stays intact when an institution or registry does this work.

Finally, FDA should recognize that all stakeholders engaged in these activities need to fund the monitoring efforts performed, whereas Section VI.B does not address this key component of establishing an effective and trustworthy monitoring paradigm. The agency should work closely with identified stakeholders to explore typical costs that enable their surveillance role and consider where responsibility lies for recovering those costs.

Q22. 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?

A22. FDA should move from regulation that effectively freezes specific model versions, prompts, or configurations toward validated performance envelopes, predetermined change control plans (PCCPs), and lifecycle oversight, all tiered by what the change does. Changes to models, prompts, retrieval, orchestration, or guardrails that remain within validated performance parameters should not automatically require a new submission. Regulation should not create an incentive to retain an older, inferior model simply because upgrading creates regulatory burden. However, it is important to note that "newer" and "scores higher overall" are different from "better." A model can improve its composite while also decreasing its precision in important clinical tasks, such as escalating a critical finding or performance for a subgroup. Evidence, rather than recency, should inform justification for reduced barriers.

There are three categories of modification: (1) changes that need no new submission under the applicable modification rules are documented and validated within the quality system; (2) changes covered by an authorized PCCP that may be implemented when they meet that plan's specified modifications, validation protocol, and acceptance criteria; and (3) changes that require additional authorization and must undergo premarket review before implementation. Passing a sponsor-defined performance threshold alone should not decide the change category. Safety-critical behaviors and relevant subgroups need their own acceptance criteria, so that aggregate improvement does not conceal unacceptable regression. The PCCP guidance already calls for sequestered test data and prespecified criteria, but we recommend independent custody of evaluation data at higher risk tiers as an added safeguard.

Additionally, the current version of the PCCP guidance does not appear to address GenAI-enabled devices explicitly. FDA should clarify the conditions under which a PCCP applies, as behavior can shift through prompts, retrieval, orchestration, guardrails, or an upstream model the sponsor does not control.

Q23. 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?

A23. Fundamentally, if a sponsor cannot predetermine the change, including specifying in advance the modification, validation protocols, and acceptance criteria, a PCCP is not the appropriate mechanism.

The current FDA guidance on PCCPs requires test data "never shown to the algorithm during training," "pre-defined acceptance criteria, and statistical tests for each planned modification," and a commitment that a modification failing evaluation "will not be implemented." It also requires modifications that are "specific, and that can be verified and validated" with an impact assessment component that rests on that specificity. To address the GenAI gap with an extension of the current architecture, FDA should consider bounded categories rather than open-ended envelopes and permit a change type with limits such as model version updates within a named family, retrieval-corpus updates from a specified source, and prompt changes within a defined function. Each change category should have corresponding acceptance criteria and an impact assessment. Then, anything outside the authorized plan must go through the applicable modification rules again to determine whether it requires a new submission.

The alternative of prespecifying the test - instead of the change - would represent a policy change that removes the specificity on which the impact assessment depends. FDA would need to clarify how an impact assessment works when no modification has been prespecified. Passing the same benchmark does not demonstrate that new failure modes have been assessed, the intended use remains unchanged, subgroup performance has been maintained, or changes to the supplier or system architecture have introduced no new hazards.

In either of the above approaches, notification would be necessary. The sponsor must notify FDA and all deployment sites whenever they modify the device to enable regulatory oversight and local governance.

Q24. 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?

A24. Continuous monitoring is important for identifying downstream impacts and mitigating harm, but it should be paired with controls that prevent unverified changes from entering clinical use. Foundation models and related services can change over time, even when presented as the same

product. Manufacturers therefore need clear visibility into which elements of the deployed configuration they control, and which may change upstream.

FDA should require appropriate supplier controls, configuration management, change detection, pre-deployment verification, and a tested fallback process. Agreements with suppliers should address behavior-affecting changes and model retirement and provide sufficient notice to complete verification. When a device can no longer maintain a validated configuration, it should transition to a validated alternative or safely suspend the affected function.

FDA can regulate the device's integrated clinical function while requiring version identification, supplier transparency, change notification, and configuration control. As the discussion paper notes, a foundation-model master file would not indicate authorization of the underlying model for a device intended use. These measures reflect established supplier-control principles rather than direct regulation of foundation models as a class.

Finally, effective local performance monitoring requires deploying organizations to have access to device outputs and relevant metadata in a structured, machine-readable format. FDA should require manufacturers to provide such access when authorization depends on local or postmarket performance monitoring.

Q25. 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?

A25. A foundation model master file (MAF) could provide FDA access to confidential component information without requiring public disclosure. However, because participation would likely be voluntary, FDA should not assume that a MAF alone will consistently provide the information needed for device review. A MAF should identify the covered model version and provide information, as appropriate, on training-data provenance, healthcare-relevant evaluation results, known limitations and failure modes, safety controls, update practices, and audit-log availability, with the level of detail proportionate to its relevance to the device's safety and effectiveness. If FDA determines that foundation model information is necessary for device review, sponsors should be required to provide that information directly or through an authorized reference to a current MAF.

Q26. 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?

A26. FDA should refine its definition of agentic AI and clarify whether the agency would intend to evaluate the characteristics identified in the discussion paper individually or in combination. FDA should assess risk based on a system's capabilities and degree of autonomy, including its ability to

take actions, access external information or tools, operate across multiple steps, and perform actions that are difficult to reverse.

FDA should also distinguish between autonomy and verifiability. If a qualified end-user cannot independently verify a system's outputs, the system should not receive credit for nominal human oversight.

As risks may arise during task execution even when the final result is correct, evaluation of agentic systems should assess both outcomes and agent actions leading to those outcomes. Testing should examine failure modes such as incorrect patient or study selection, duplicate actions, stale or incomplete data, failed handoffs, partial task execution, and variability across repeated runs of the same task. Sponsors should evaluate repeated runs of the same task and report meaningful variability in outcomes and intermediate actions rather than solely final outputs.

Agentic systems may introduce cybersecurity and prompt-injection risks through user inputs, images, retrieved content, tool outputs, and other external data sources. FDA should align agentic-AI expectations with existing cybersecurity requirements, including testing of these attack pathways and application of least-privilege principles for tool access and external actions.

Controls should be proportionate to the consequences and reversibility of the actions a system can perform. Drafting an order, submitting it, and submitting it irreversibly present different risks and may warrant different levels of validation and oversight. Higher-risk functions may warrant clinically representative validation, phased deployment, postmarket monitoring, and reversibility analysis. Where human supervision is a claimed safeguard, FDA should evaluate that interaction directly and define the supervisor's role.

As always, the ACR welcomes continued dialogue with the FDA on these and other radiology-relevant topics. For more information, please contact Michael Peters, Senior Director, ACR Government Affairs, at mpeters@acr.org.

Sincerely,



Dana H. Smetherman, MD, MPH, MBA, FACR
Chief Executive Officer