HEALTH INNOVATION ALLIANCE

October 19, 2026

Food and Drug Administration
Center for Devices and Radiological Health
Digital Health Center of Excellence
10903 New Hampshire Avenue
Silver Spring, MD 20993

**Re: Considerations for the Regulation of Generative AI-Enabled Medical Devices: Discussion Paper and Request for Feedback (Docket No. FDA-2026-N-7874)**

*Submitted electronically via regulations.gov*

Dear FDA Digital Health Center of Excellence Team,

HIA is a coalition of patient advocates, specialty care providers, and technology companies that has advanced a risk-based, innovation-forward approach to health technology policy for nearly two decades. Our response below builds on HIA's February 2026 comments on HHS's Health Sector AI RFI[i], HIA's earlier comments on FDA's AI-enabled device performance RFI[ii], HIA's Clinical AI Confidence Compact concept, which was shared directly with FDA staff before this discussion paper's publication, and conversations with patient advocacy and provider organizations over the summer of 2026, none of which called for a pause or new regulatory oversight of clinical AI.

Autonomous clinical AI is not a future scenario for FDA to regulate cautiously into existence. It is here, and it is needed. What's missing is clarity and certainty on whether and how oversight makes sense. Our core message is simple: accelerate real-world evidence generation.

Independent studies in imaging, primary care, and cardiology show AI, alone or alongside clinicians, meeting or exceeding clinician-only benchmarks on specific, well-defined tasks.[iii] Independent evaluators also caution that measurable, real-world outcome evidence remains thin relative to model capability.[iv] HIA reads that gap as the case for phased, monitored, time-limited deployment that generates the missing evidence.

The real risk to patients is unsubstantiated risk itself. Treating every use case as equally uncertain and equally dangerous, without weighing alternatives such as no care at all, does not protect patients. It compounds an access problem this technology is positioned to help solve. The United States faces a looming physician shortage.[v] Ninety-two percent of rural counties are federally designated primary care shortage areas.[vi] For a patient in one of those counties, the realistic comparator for an AI-delivered diagnosis is not “a clinician, or an AI” — it is “an AI, or nothing.”

A second, more immediate risk is a state-by-state patchwork of AI-in-health-care legislation, much of it moving faster than this federal process, that threatens to create fifty different, cumbersome pathways for bringing the same innovation to market. Federal action can lessen fragmentation by giving states a credible, evidence-generating federal alternative to enacting rules alone.

Industry's own posture confirms the cost of continued ambiguity. Most generative AI tools on the market today are, in one attorney's description, "dancing around the FDA-regulated space" — deliberately avoiding device-triggering functions until FDA clarifies what will actually be required, because, in her words, "[n]o one really wants to be the guinea pig."[vii] Companies that have already accepted FDA regulation, the same attorney adds, are looking for more clarity, and at least one company CEO describes regulatory leadership from FDA as the most important safeguard against either a patchwork of different regulations or bad actors and reactionary responses.[viii] Ambiguity limits innovative and safe products from reaching patients.

HIA welcomes this discussion paper and supports its general framing, in particular:

- Its two-axis risk framework, with two refinements: the activity axis should measure reliance (whose judgment an outcome depends on), and the consequence axis should reflect the chance, frequency, and severity of an inaccurate output.
- Its openness to accepting greater premarket uncertainty in exchange for real-world postmarket monitoring.
- Its competency-based approach, benchmarking paired with clinical confirmation, modeled on how physicians are trained and credentialed, and its recognition that clinical confirmation need not always require a prospective study.

We do caution that the framing of this paper could be read to reach software functions outside the device definition. In particular, FDA's framework should:

- Continue to distinguish AI-enabled tools that support health care professionals' independent decision-making from software that performs regulated medical device functions.
- Reference the non-device CDS criteria that the 21st Century Cures Act added to the FD&C Act at section 520(o)(1)(E).

A discussion paper is not a policy. HIA is asking FDA to act now: build on TEMPO to create a Clinical AI Confidence Compact ("Compact") with interested U.S. states and companies to generate the evidence this paper says it needs, while patients are already being served.

HIA's proposed Compact is a time-limited agreement under which a defined AI-delivered care model operates with phased autonomy (full clinician review, then sampling review, then monitoring) and continuous reporting, to generate real-world evidence on where risk actually lies. Such Compacts would let real-world evidence, generated under a phased, monitored, and temporary deployment, count as valid clinical confirmation evidence under the framework this paper proposes.

## Detailed Response

The responses below follow the discussion paper's own section structure. HIA's principal recommendations are in Questions 1, 2, 11, 15, 18, and 22 and in Part E.

### A. Section IV — Risk Assessment Framework

***Question 1 — Does the two-axis framework capture the right dimensions?***

HIA agrees the two-axis framework — activity and consequence — captures the dimensions that matter most. We would caution against adding dimensions such as reversibility or traceability as independent axes. They should be treated as modifiers within the existing framework.

HIA would clarify what the activity axis is actually measuring. Read narrowly, "activity" can seem to describe what the AI does. What matters clinically is different: as a given diagnosis, treatment course, or clinical finding is reached, how much does it depend on an informed person's judgment, versus the AI operating on its own. That reliance is what should place a function on this axis (see more, Question 2).

HIA also recommends clarifying the consequence axis. HIA has previously defined health IT risk as a function of three factors: the chance that a system's output will be inaccurate, the frequency of potential inaccuracy, and the severity of the outcome if an inaccurate output is relied upon, including whether the harm is reversible.[ix] HIA urges FDA to treat consequence as these three components.

Severity sets a reference point between clinical domains: an inaccurate insulin dose carries more severe potential consequences than an inaccurate poison-ivy recommendation, as a category. That comparison does not mean every instance of autonomous insulin dosing carries the same risk. Within a single clinical domain, actual risk still varies with the chance and frequency of inaccuracy. Benchmarking, real-world evidence, and monitoring affect those risk factors. A function's consequence tier should therefore be assessed using real-world data on its actual chance and frequency of error, including missed escalation triggers that were not anticipated. All three components should also be judged against the realistic alternative available to the patient (see more, Question 15).

***Question 2 — How should the non-directive to action-directing spectrum be assessed?***

Consistent with Question 1 above, directiveness exists on a continuum. Wording alone ("recommend," "should") is a poor proxy for where a function sits on it. Time pressure is also an uncertain proxy; the automation-bias literature reports mixed effects. The variable that matters is reliance: as a given diagnosis, treatment course, or clinical finding was reached, how much did it depend on a person's judgment, versus the AI operating on its own.

Congress already used reliance as the organizing idea in the CDS exclusion that the 21st Century Cures Act added to the FD&C Act (section 520(o)(1)(E)). HIA's point is that reliance is better measured directly by asking whose judgment produced the outcome.

Reliance is a property of a function's design and intended use. Escalation triggers or pathways for clinician review and the system's containment capabilities, including the ability to withdraw autonomy rapidly, shift reliance from the AI function to an HCP. Consider an AI function for uncomplicated urinary tract infection, one of the low-complexity conditions HIA's Clinical AI Confidence Compact identifies for Track 1. A version of this function that routes red-flag presentations to a clinician sits closer to the clinician-reliant end of the axis than an otherwise identical function with no such trigger.

In short, a function's position on the axis is set by reliance, as assessed from its design (including its escalation pathways) and the evidence on how it performs in use. That evidence can support a defined change to the function's design and scope, such as narrowing or widening the conditions under which autonomy is withdrawn, through the PCCP mechanism discussed under Question 22.

***Question 3 — Do patient-facing functions present different or higher risk than HCP-facing functions?***

Yes, and consistent with Questions 1 and 2: a patient-facing function could sit higher on the activity axis, depending on the user's reliance on the AI system versus a clinician. Yet, the relevant comparison for a patient with no HCP in the loop is AI versus whatever the patient would rely on otherwise: unmoderated information, delayed care, or no care at all (Question 15).

Also, its consequence assessment should still be evaluated on the three dimensions above: chance, frequency, and severity of an inaccurate output, including reversibility.

***Question 6 — Weighing under- vs. over-escalation error***

Both failure modes should be weighed together. Over-escalation carries its own cost: unnecessary emergency-department utilization, patient anxiety, and, over time, erosion of trust that pushes patients back toward no care at all. In an access-constrained system, directing a patient toward care they cannot reach or afford has much the same effect as under-escalation. Benchmarking for escalation functions should measure both directions against defined thresholds.

**B. Section V — Competency-Based Premarket Evaluation**

FDA grounds this competency-based approach in how physicians themselves are evaluated and credentialed, a construct similar to HIA's Clinical AI Confidence Compact.

Device benchmarking is necessary, but insufficient alone to demonstrate real-world competence. Clinical confirmation is where independence expands as competence is demonstrated. The Compact's phases track this directly with supervision narrowing as performance holds up. At each stage, the question is the one raised in Question 2 above: on whose judgment the clinical outcome actually depends.

***Question 7 — Is the competency-based approach (benchmarking plus clinical confirmation) sound?***

Yes. The two-part structure fits GenAI's open-ended behavior better than exhaustive input/output testing, for the reasons above. FDA's Section V.C methods — shadow deployment, clinician adjudication, prospective study — already describe most of the phased structure in HIA's Compact concept. See Question 11.

***Question 9 — Are there missing or redundant benchmarking elements?***

HIA recommends one addition: an explicit standard-of-care/guideline-concordance element, alongside the ground-truth-based elements already listed, measuring not only whether an output is technically correct but whether it matches or exceeds what a qualified clinician panel, or a recognized clinical-guideline source, would recommend for the same patient in the same circumstances. This is the comparator question raised in Section V.D.1 and deserves its own benchmarking element.

***Questions 10 and 16 — Benchmark validity and independent third parties***

Benchmarks created by device developers themselves should remain an acceptable component of competency-based evaluation, provided their methodology is transparent and their clinical relevance is validated against established clinical guidelines and representative clinical scenarios. Depending on the function's risk profile, validation may include independent qualified clinician review, established clinical-guideline and evidence platforms, or other appropriate independent corroboration. Limiting evaluation to third-party or FDA-developed benchmarks alone could concentrate evaluation capacity in a small number of institutions and slow the field where CDRH states it wants a scalable, high-throughput approach. For the same reason, third-party evaluation should be an option where warranted.

***Question 11 — Should prospective study be required in every case?***

It should not. FDA's named methods — shadow deployment, clinician adjudication, prospective study — describe increasing levels of rigor that a real-world deployment structure like HIA's Compact is designed to move through as evidence accrues. HIA urges FDA to state plainly that evidence generated in this way satisfies Section V.C clinical confirmation.

***Question 14 — How should comparators and acceptance criteria be selected for open-ended outputs?***

Comparing performance to a panel of qualified clinicians reflecting the standard of care is sound. FDA should not treat performance equal to a median clinician as the ceiling of an appropriate comparator standard. For narrowly defined, high-volume, well-evidenced tasks, the right comparator may be the performance of expert, seasoned specialists, because for those specific tasks the evidence shows some systems already clear it. HIA also cautions against defaulting to specialist-panel comparators as uniformly more rigorous than generalist comparators; the right standard depends on the population the device is intended to serve and the alternative realistically available to it, addressed further under Question 15.

On human-AI team versus fully autonomous comparison: HIA urges FDA not to treat human-AI team performance as inherently the safer or preferred standard by default. A pooled analysis of 52 studies found human-AI teams beat the clinician alone in 95 percent of conditions but beat the AI alone in only 44 percent.[x] Others reach similar conclusions.[xi]

Autonomous operation is not uniformly safer regardless of task: a 2026 safety benchmark found top models produced potentially severely harmful recommendations in roughly one in fourteen consultations, though they outperformed generalist physicians on the same cases.[xii] This is why competency-based benchmarking, paired with phased, monitored deployment, is the right tool for managing this risk. Requiring human review as a categorical safety measure, without evidence that the specific human-AI configuration improves outcomes, risks codifying a comparator that sounds safer than it measurably is.

***Question 15 — Should performance be assessed relative to the care likely to occur in the device's absence?***

HIA urges FDA to formally recognize no clinical encounter at all as a common and important comparator for real-world deployment, particularly for populations with limited access to alternatives.[xiii] For a substantial share of the population these tools are intended to serve, the realistic choice is AI or nothing. A framework that measures AI function only against clinician performance, without also measuring what happens when a comparably situated patient has no functioning access to a clinician at all, will understate this technology's benefit where that benefit could be largest.

**C. Section VI — Postmarket Monitoring**

***Question 18 — Should FDA accept greater premarket uncertainty through greater reliance on postmarket monitoring?***

Yes, greater reliance on postmarket monitoring makes sense, in proportion to risk. CDRH's own examples already draw this line, contrasting autonomous antibiotic prescribing for confirmed strep throat with autonomous thrombolytic initiation in a stroke workflow. Strep throat sits squarely in the same category as HIA's Clinical AI Confidence Compact Track 1 conditions: well-established, evidence-based protocols; low clinical complexity; episodic conditions such as uncomplicated UTI, acute sinusitis, or bacterial conjunctivitis. Where an established protocol lowers the chance of inaccuracy, the condition's severity is low even if mismanaged, and escalation triggers bound the frequency of unmitigated error, FDA should accept meaningfully reduced premarket evidence in exchange for robust, real-time postmarket monitoring. The appropriate balance should reflect the clinical risk of the intended function, the strength of existing evidence, and the ability of monitoring and escalation safeguards to identify and mitigate potential harm.

***Question 19 — Periodic re-benchmarking, sample-based review, and drift monitoring***

HIA supports these as the right building blocks. They are how the chance-and-frequency evidence described under Question 1 gets generated. HIA's Clinical AI Confidence Compact concept explores a similar structure on a temporary basis, regular reporting on volume, escalation rates, safety events, and complaints, with a defined transition from full clinician review to sampling to steady-state monitoring, as a way to generate evidence on whether and where risk actually lies.

***Question 20 — Machine-based supervisory agents***

This is a promising area for continued exploration. HIA would support a future, specifically scoped FDA effort here once the Section V framework matures.

***Question 21 — Roles for clinicians and institutions without diffusing manufacturer accountability***

HIA recommends a sliding-scale approach. The manufacturer's postmarket monitoring and real-time escalation obligations should scale with the function's activity and consequence tier, so accountability remains with the manufacturer by default and is only shared where a treating clinician or institution is actually present. FDA's own interoperable automated insulin delivery framework already reflects this logic, holding manufacturers accountable for real-time, autonomous dosing decisions made with no clinician in the loop at the time of use, allocated by each component's function.[xiv] This connects to the liability question raised in E.1 below, which this discussion paper does not address.

***Question 22 — Premarket assessment as a baseline for postmarket modifications***

HIA has separately and repeatedly urged FDA to unlock the PCCP pathway's full potential. Using the premarket competency-based assessment as a baseline against which modifications are re-evaluated is the right model. HIA urges FDA to pair this discussion paper's framework with a commitment to approve broader, pre-specified evolutionary pathways for GenAI-enabled devices specifically, given how frequently these systems are expected to change relative to traditional software.

***Question 23 — PCCP when future modifications cannot be fully prespecified***

HIA would support a bounded evolutionary envelope approach: prespecifying categories and limits of change, such as retraining cadence or permitted guardrail modifications. Requiring every specific future change to be named in advance is unworkable.

**D. Section VII — Other Topics**

***Question 25 — Foundation Model Device Master Files***

HIA supports the voluntary Foundation Model MAF concept in principle, as a way to give FDA reference information without a new mandatory disclosure burden on device sponsors who rely on third-party models. Its value will depend on whether model developers see sufficient benefit, such as faster review for devices built on their models, to participate voluntarily.

***Question 26 — Additional considerations for agentic AI systems***

The agentic AI competencies FDA proposes under element A.1 are a reasonable start, consistent with how the field is now defining this category.[xv] HIA notes that Track 2 of its Compact concept, continuous monitoring with defined escalation triggers for complex, ongoing conditions, is directly relevant to agentic systems that plan and execute multi-step actions, and welcomes the opportunity to discuss how Compact-style monitoring could complement agentic-specific premarket benchmarking.

**E. Issues This Discussion Paper Does Not Ask About, But Should**

***1. Liability and safe harbor***

This discussion paper is silent on liability. HIA has proposed a safe harbor model tied to substantial compliance with a monitoring structure like the Compact: a participant in substantial compliance with phased-autonomy monitoring and reporting obligations should have protection against liability claims except in cases of fraud, gross negligence, or willful misconduct, while patients who decline AI-delivered care retain full recourse under ordinary standards through the non-AI alternative that must remain available to them. HIA urges FDA to engage on how its evidentiary framework could inform, or be informed by, parallel HHS and state work on liability standards for compliant deployments.

***2. State practice-of-medicine authority***

FDA authorization of a device software function does not resolve whether that function may lawfully deliver care without a licensed clinician under a state's practice-of-medicine laws. The resulting state-by-state patchwork is a significant barrier to clinical AI deployment. HIA urges FDA to support a mechanism, such as the tri-party structure in HIA's Compact, that pairs FDA's enforcement discretion with state practice-of-medicine relief under a single time-bound agreement, and to use what it learns to inform any longer-term policy fix.

***3. Scope of the framework***

FDA's paper addresses software functions that meet the device definition. Section 520(o)(1)(E) of the FD&C Act, added by the 21st Century Cures Act, excludes from that definition clinical decision support that supports a health care professional's judgment and enables the professional to independently review the basis for its recommendations, so that the professional does not rely primarily on them. HIA asks FDA to state in any follow-on guidance that the risk framework, including its treatment of action-directing functions, applies only to functions that meet the device definition and does not alter the 520(o)(1)(E) criteria.

**Conclusion**

HIA appreciates FDA's engagement on this topic and shares its view that a risk-based, proportionate approach is right. We ask FDA to commit to recognizing real-world evidence generated under phased, monitored, time-limited Compact deployments as valid clinical confirmation evidence under Section V.C, and, through TEMPO or a similar structure, to identify a cohort of low-complexity, well-established-protocol functions consistent with Track 1 for a near-term pilot. HIA and its members stand ready to help.

Sincerely,



**Roslyn Docktor**
Executive Director
Health Innovation Alliance

[i] HIA_AI_RFI_Response_2.23.26.pdf
[ii] HIA-Comments-FDA-AI-Enabled-Device-Performance-RFI-vFIN-1.pdf
[iii] Eisemann N, Bunk S, Mukama T, et al. Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. Nat Med. 2025;31:917–924. https://www.nature.com/articles/s41591-024-03408-6; Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digit Med. 2018;1:39. https://www.nature.com/articles/s41746-018-0040-6; see also

Huang JJ, et al. Autonomous artificial intelligence for diabetic eye disease increases access and health equity in underserved populations. npj Digit Med. 2024;7:196. https://www.nature.com/articles/s41746-024-01197-3; Korom R, et al. Real-world deployment of a generative AI clinical decision support tool across primary care clinics in Kenya (39,849 patient visits), as reported in ARISE Network, The 2026 Healthcare AI Industry Report (July 2026), at 34, https://www.arise-ai.org/reports/2026-healthcare-ai-industry-report.pdf; Zhang, et al. Clinician-artificial intelligence collaboration: a win-win solution for efficiency and reliability in atrial fibrillation diagnosis (2025), as reported in id. at 39.

[iv] ARISE Network (Perez A, Tusty M, Liu C, et al.). The 2026 Healthcare AI Industry Report. Stanford, CA; July 2026. https://www.arise-ai.org/reports/2026-healthcare-ai-industry-report.pdf

[v] American Hospital Association. Fact Sheet: Telehealth. Feb. 2025 (reporting the Association of American Medical Colleges' projected shortage of up to 86,000 physicians by 2036). https://www.aha.org/fact-sheets/2025-02-07-fact-sheet-telehealth

[vi] Horstman C, Shah A. The State of Rural Primary Care in the United States. The Commonwealth Fund; Nov. 17, 2025 (reporting that 92 percent of rural counties were designated primary care health professional shortage areas in 2023). https://www.commonwealthfund.org/publications/issue-briefs/2025/nov/state-rural-primary-care-united-states

[vii] Reuter E. 4 questions about the FDA's approach to generative AI. MedTech Dive; Sept. 9, 2026 (quoting Suzanne Levy Friedman, Partner, Honigman). https://www.medtechdive.com/news/4-questions-about-the-fdas-approach-to-generative-ai/829871/

[viii] Reuter, supra note vii (paraphrasing Sharif Vakili, CEO, UpDoc).

[ix] Health Innovation Alliance. Use Cases for AI Tools to Improve Health Care. March 2025, at 7 (defining health IT risk); see also id. at 10 (listing level of reliance on AI output for decision or action among the factors influencing risk). https://health-innovation.org/wp-content/uploads/2025/03/HIA-AI-Use-Cases-Report-2025.pdf

[x] Liu Y, et al. Pooled reliability analysis of human-AI team performance across 52 studies and 87 clinical conditions (2025), as reported in ARISE Network, supra note iv, at 41.

[xi] Emanuel EJ, Baker-Butler A, Khosla N, Khosla V. Will autonomous AI exceed AI-aided physicians as the best medical care? JAMA. 2026;336(11):915–918. https://doi.org/10.1001/jama.2026.15380; Warren LM, et al. Retrospective evaluation of AI as a second reader in NHS breast cancer screening, including arbitration outcomes (2026), as reported in ARISE Network, supra note iv, at 42.

[xii] Wu D, et al. NOHARM: a benchmark for clinically harmful model outputs (2026), as reported in ARISE Network, supra note iv, at 26–27.

[xiii] See supra notes v and vi; see also Samuels J. America's least-discussed physician shortage is getting worse. Forbes; Aug. 4, 2026 (reporting access gaps in geriatrics, sleep medicine, and family medicine). https://www.forbes.com/sites/johnsamuels/2026/08/04/americas-least-discussed-physician-shortage-is-getting-worse/

[xiv] See 21 C.F.R. § 862.1356 (interoperable automated glycemic controller; Class II, special controls), https://www.ecfr.gov/current/title-21/chapter-I/subchapter-H/part-862/subpart-B/section-862.1356; see also 21 C.F.R. § 862.1355 (integrated continuous glucose monitoring system); 21 C.F.R. § 880.5730 (alternate controller enabled infusion pump).

[xv] Angus DC, Khera R, Lieu T, et al. AI, Health, and Health Care Today and Tomorrow: The JAMA Summit Report on Artificial Intelligence. JAMA. 2025;334(18):1650–1664. https://doi.org/10.1001/jama.2025.18490