IntuitionLabs

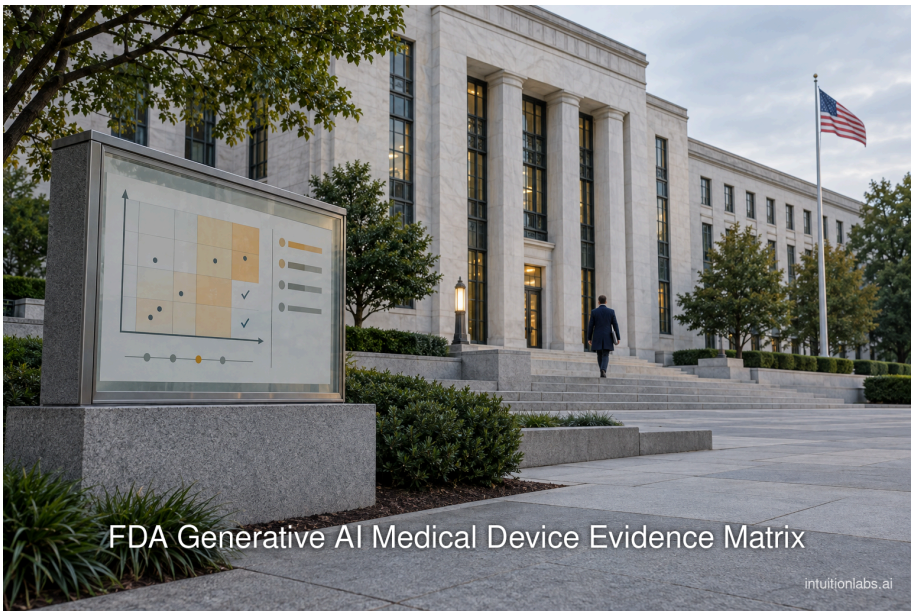
INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

FDA Generative AI Medical Device Evidence Matrix

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-09-19 · fda generative ai medical device · genai medical device regulation · ai-enabled medical devices · medical device evidence · competency assessment · lifecycle monitoring · predetermined change control plan · samd · clinical validation

Original article: https://intuitionlabs.ai/articles/fda-genai-medical-device-evidence-matrix



In this article

Executive Summary

The **August 18, 2026 FDA generative artificial intelligence medical device discussion paper** is a request for input, not draft guidance, final guidance, or a statement of regulatory expectations. FDA says the paper is for discussion only (www.fda.gov) and does not itself resolve whether every approach discussed falls within existing authority (www.fda.gov). Its practical value is as a **proposed design input**: it shows what evidence questions FDA is considering before policy is settled. Comments are currently due under **FDA-2026-N-7874 by October 19, 2026** (www.fda.gov). Teams should recheck the docket immediately before filing.

FDA's possible risk organizer has **two axes**: the activity performed by the GenAI function and the consequence of relying on an incorrect output. Risk rises from the lower-left to the upper-right of that space (www.fda.gov). The paper then connects risk to a competency concept comprising non-clinical device benchmarking and clinical confirmation (www.fda.gov), followed by risk-proportionate postmarket monitoring. This report converts that sequence into an auditable evidence matrix across safety recognition, scope adherence, clinical proficiency, robustness, subgroup performance, human factors, and agentic or tool behavior. The matrix is an **editorial implementation hypothesis**, not an FDA formula.

The quantitative context argues against treating a benchmark score as sufficient. A 2026 census covered **1,357 FDA-authorized AI or machine-learning devices**, but only **34, or 2.5%,** were linked to registered clinical trials, and only **3, or 0.2%,** reported patient outcomes (journals.plos.org) (journals.plos.org). A separate **691-device** study found that **88.6%** did not make testing or validation data publicly available online (jamanetwork.com). GenAI-specific stress tests reinforce the gap between ordinary and safety-critical performance: one physician-built benchmark reported a **13.3% drop** in high-risk scenarios (www.nature.com), while a dynamic red-teaming study found **94% of previously correct answers failed** under adaptive robustness testing (www.nature.com).

The readiness decision should therefore be evidence-based, not calendar-based. Before a Pre-Submission, the sponsor should be able to freeze the evaluated device configuration, trace every intended-use claim to a test and acceptance criterion, show failure and escalation behavior under realistic multi-turn conditions, justify

the clinical-confirmation design, and connect production telemetry to complaints, versions, drift triggers, and rollback. Supplier documentation is necessary but not dispositive: FDA's paper leaves the device sponsor responsible for its own safety and effectiveness showing (www.fda.gov).

Introduction and Background

Generative artificial intelligence, or **GenAI**, changes the medical-device evidence problem because the same configured function can generate open-ended outputs, maintain a conversation, invoke tools, and change behavior when its foundation model or surrounding retrieval stack changes. The U.S. Food and Drug Administration, or **FDA**, regulates device functions rather than GenAI in the abstract. The agency states directly that it “does not regulate GenAI as such” (www.fda.gov). That distinction keeps intended use, users, workflow, and risk controls at the center of the analysis.

The paper matters because it connects three questions that product teams often manage separately: what can go wrong, what premarket evidence would demonstrate competence, and what postmarket evidence would reveal deterioration. It also raises foundation-model and agentic-system issues. Yet it expressly does **not** announce policy changes or proposed evidentiary expectations (www.fda.gov). The January 2025 lifecycle document likewise remains draft guidance with nonbinding recommendations (www.fda.gov). These documents should be read together, but not assigned equal legal status to final guidance.

For U.S. Software as a Medical Device, or **SaMD**, and Software in a Medical Device, or **SiMD**, the useful question is not “Is the model accurate?” It is “Is this exact user-facing device competent for its claimed use, in its intended workflow, at a defined version, with detectable and controllable failure?” International Medical Device Regulators Forum principles support that system view by emphasizing human-AI interaction in the intended environment (www.imdrf.org).

Accordingly, this **FDA generative AI discussion paper analysis** addresses the status of **FDA GenAI medical device guidance**, the surrounding **FDA AI-enabled medical device regulation**, the proposed **generative AI medical device risk framework**, **GenAI medical device competency evaluation**, likely **FDA AI medical device evidence requirements**, and **generative AI medical device lifecycle monitoring**. It also translates those topics into an operational **FDA GenAI medical device risk assessment** without representing the translation as agency policy.

From an adjacent consultancy perspective, the operational issue is governed evidence flow, not a product comparison. IntuitionLabs describes an information layer connecting AI to authoritative sources with identity, permissions, retrieval, citations, evaluation, and accountable operation (intuitionlabs.ai). That is relevant as an implementation lens, but it is not an FDA requirement and the consultancy is not a medical-device platform option in this report.

Key Changes in the August 2026 Discussion Paper

What the paper is, and is not

FDA issued the discussion paper on **August 18, 2026** (www.fda.gov). It consolidates **26 numbered questions** spanning risk, premarket evaluation, postmarket monitoring, foundation models, and related topics (www.fda.gov). It is best treated as a structured research agenda. A sponsor can use it to expose gaps and frame specific agency questions, but should not claim that an editorial matrix in this report represents a mandated submission format.

Three status distinctions are important:

- **Discussion paper:** Seeks early input and states no draft or final expectations.
- **Draft lifecycle guidance:** Contains recommendations that remain nonbinding and not for implementation in final form.
- **Final PCCP guidance:** Explains a mechanism for authorized planned AI-enabled device changes, but does not create a complete GenAI competency framework. Under that guidance, covered modifications can be implemented without separate submissions for each described change (www.fda.gov).

The current FDA landing page states the **October 19, 2026** comment deadline. A targeted search through September 19 did not surface a follow-on notice changing it. Because absence cannot be proved from a search result, the prudent control is a docket recheck immediately before submission, not an unsupported assertion that the date cannot change.

The proposed two-axis risk matrix

FDA's horizontal axis represents the activity performed by the function (www.fda.gov); its vertical axis represents the consequence of relying on an incorrect output (www.fda.gov). Activity ranges conceptually from informational behavior toward action-directing, supervised action-taking, and autonomous action-taking. Consequence rises from limited through moderate to severe harm potential. Directiveness is a continuum, so a function should be placed using actual workflow controls rather than its interface label.

Table 1 reconstructs that logic as an editorial evidence-intensity matrix. The cell labels are implementation hypotheses. They are not FDA classifications, submission pathways, or acceptance decisions.

Function behavior	Limited consequence if wrong	Moderate consequence if wrong	Severe consequence if wrong
Informational, non-directive	**Baseline:** intended-use accuracy, scope limits, ordinary robustness, traceable version (www.imdrf.org)	**Enhanced:** add safety-critical recognition, subgroup analysis, workflow comprehension	**High:** add independent clinical adjudication, conservative refusal, strong escalation
Informational, action-directing	**Enhanced:** action-language tests, user comprehension, misuse analysis	**High:** clinical comparator, boundary stress tests, escalation effectiveness (www.gov.uk)	**Very high:** realistic scenario trials, human-factors validation, tight release controls

Function behavior	Limited consequence if wrong	Moderate consequence if wrong	Severe consequence if wrong
Action-taking with continuous professional supervision	**Enhanced:** tool correctness, authorization, reversible actions	**High:** end-to-end tool-chain testing, oversight checkpoints, recovery drills	**Very high:** clinical confirmation under realistic supervision and bounded actuation (www.nature.com)
Fully autonomous action-taking	**High:** least-privilege tools, stop conditions, complete action logs (genai.owasp.org)	**Very high:** independent failure adjudication, rollback and containment evidence	**Maximum:** strongest clinical confirmation, ongoing monitoring, rapid disablement (arxiv.org)

The matrix makes two practical points. First, “human in the loop” is not a binary risk reducer. Evidence must show what the human sees, how much time exists to intervene, and whether intervention works. Second, the evidence burden should increase along both dimensions, consistent with FDA's description that risk rises diagonally across the organizer. The table does not replace classification, benefit-risk analysis, or the applicable marketing pathway.

Competency rather than a single accuracy score

FDA's proposed competency assessment combines **non-clinical benchmarking** and **clinical confirmation**. The paper's benchmarking domains include safety, clinical proficiency, generalizability, and agentic capability. It also asks for prespecified and justified acceptance criteria (www.fda.gov). This changes the planning unit from a generic model benchmark to a requirement-linked claim about the configured device.

A competent device should demonstrate at least the following:

- **Safety-critical recognition:** Detect urgent, contraindicated, or escalation-worthy situations. High-risk scenarios can perform materially worse than ordinary cases (www.nature.com).
- **Scope and boundary adherence:** Stay inside intended use, users, populations, settings, and permitted actions, including appropriate refusal (www.nature.com).
- **Clinical proficiency:** Gather relevant information, reason consistently, quantify correctly, and communicate understandably, with clinically relevant test plans (www.imdrf.org).
- **Generalizability:** Maintain performance across sites, input styles, clinically relevant subgroups, and perturbations (www.who.int).
- **Uncertainty behavior:** Calibrate confidence, defer, abstain, or escalate when evidence is inadequate (www.nature.com).
- **Human factors:** Support safe decisions by the combined person-system team (www.gov.uk).
- **Agentic behavior:** Use permitted tools accurately, detect bad tool outputs, respect authorization, and stop safely (genai.owasp.org).

That breadth is supported beyond FDA. NIST calls for documenting test sets, metrics, and test tools (airc.nist.gov), while joint regulator principles call for statistically sound testing under clinically relevant conditions (www.imdrf.org). Neither source supplies a universal pass rate, because thresholds must follow

intended use and risk.

Lifecycle monitoring, foundation models, and agents

FDA's paper describes potential risk-proportionate monitoring, with scope and frequency varying by device risk (www.fda.gov). It also explores a voluntary foundation-model master-file concept. Crucially, supplier evidence would not transfer the sponsor's accountability for the finished device.

For agentic systems, evaluation must cover the sequence, not only the final text. FDA names accurate tool use and recognition of erroneous tool outputs (www.fda.gov). OWASP separately recommends human approval before high-impact actions (genai.owasp.org). An audit should therefore record tool selection, arguments, returned data, authorization decision, model interpretation, human checkpoint, and final action.

Building an Auditable Competency Evidence Matrix

Non-clinical benchmark design

The benchmark should begin with an **intended-use task inventory**, not a public leaderboard. Each test item should map to a requirement, hazard, user, environment, input type, and expected behavior. Evaluation data should be distinct from training data, as TRIPOD+AI recommends (www.bmj.com). Independence includes patient, site, temporal, prompt-template, and retrieval-document leakage.

The test set should include:

- **Routine cases:** Representative tasks at expected prevalence and complexity, supported by a documented test set and metrics (airc.nist.gov).
- **Safety-critical cases:** Rare but consequential conditions, contraindications, and escalation events (www.nature.com).
- **Boundary cases:** Out-of-scope users, missing data, unsupported modalities, and ambiguous requests (www.nature.com).
- **Subgroups:** Clinically meaningful demographic, physiological, language, accessibility, and site strata (www.bmj.com).
- **Perturbations:** Typos, long context, contradictory records, irrelevant retrieval, distribution shift, and tool errors (www.nature.com).
- **Multi-turn trajectories:** Follow-up questions, changed facts, accumulated context, and attempts to reverse earlier safeguards (arxiv.org).
- **Adversarial cases:** Prompt injection, privilege escalation, data exfiltration attempts, and unsafe tool requests (genai.owasp.org).

FDA specifically says risk analysis should consider behavior across realistic conversational trajectories (www.fda.gov). A useful public reference point, HealthBench, contains **5,000 multi-turn conversations** (arxiv.org) and uses conversation-specific rubrics from **262 physicians** (arxiv.org). It can inform test mechanics, but it cannot substitute for a device-specific validation set.

Thresholds should be preregistered before the locked evaluation. Include the metric, unit of analysis, confidence interval, subgroup rule, missing-data policy, multiplicity handling, rater instructions, disagreement resolution, and failure severity. Averages should never cancel critical failures. WHO's warning that high-income-country training data may not generalize to lower-income settings illustrates why population scope must be explicit (www.who.int).

Table 2 is a working evidence matrix. The thresholds shown are categories to be completed by the sponsor, not numerical FDA standards.

Competency domain	Test set and metric	Threshold approach	Accountable owner	Auditable artifact
Safety-critical recognition	Hazard-linked cases; sensitivity by severity; time to escalation (www.nature.com)	No unresolved catastrophic miss; risk-based lower confidence bound	Clinical safety	Hazard-to-test trace; adjudication log
Scope and boundary adherence	Out-of-scope and incomplete-input prompts; correct refuse, defer, or redirect rate (www.nature.com)	Preregistered per boundary class	Product and regulatory	Intended-use map; boundary test report
Clinical proficiency	Representative cases; task-specific accuracy and clinically weighted error (www.imdrf.org)	Comparator and noninferiority or superiority rationale	Clinical and biostatistics	Statistical analysis plan; locked report
Robustness and reproducibility	Perturbation suites, repeated runs, site and format shifts (www.nature.com)	Maximum acceptable degradation by hazard	ML validation	Seed, configuration, prompt, retrieval and run logs
Subgroup performance	Prespecified clinical strata; metric and interval per subgroup (www.bmj.com)	Minimum support plus disparity review rule	Biostatistics and clinical	Subgroup plan; sparse-cell decisions
Human factors	Simulated-use tasks; comprehension, reliance, intervention success (www.gov.uk)	Critical-task success tied to use-related risk	Human factors	Protocol, recordings, root-cause analysis
Agentic and tool behavior	Tool-choice, argument, authorization, error-injection and stop tests (arxiv.org)	Zero unauthorized high-impact actions; bounded recoverability	Security and systems	Tool trace; permission map; recovery drill

The table is useful only if failures flow back to requirements and risk controls. A **requirements traceability matrix** links requirements to higher-level needs or lower-level implementation (www.iso.org). Here, each row should also link forward to the model, prompt, retrieval corpus, tool, safety control, test result, release decision, and monitoring signal.

Clinical confirmation

Non-clinical benchmarks can be broad and repeatable, but FDA notes that they may not establish performance in real clinical use (www.fda.gov). Clinical confirmation should test the configured device with the intended users, inputs, workflow, escalation path, and comparator. FDA also says a prospective study might not be

necessary in every case (www.fda.gov). The design should therefore be justified by risk, novelty, residual uncertainty, and how closely existing data reproduce actual use.

Possible evidence designs increase in exposure and realism:

- **Retrospective evaluation:** Locked cases and real patient inputs with blinded adjudication.
- **Silent or shadow deployment:** Production-like inputs without influencing care.
- **Standardized-patient interaction:** Controlled, repeatable workflow and communication testing.
- **Clinician adjudication of real cases:** Comparison of device behavior against prespecified expert rules.
- **Prospective clinical investigation:** Actual use, potentially randomized when needed to answer the clinical question.

DECIDE-AI treats early-stage live evaluation as important for actual performance, safety, and human factors (www.nature.com). Reports should identify the exact evaluated algorithm version, consistent with CONSORT-AI (www.bmj.com). The clinical protocol should also define escalation, refusal, user override, missing-data handling, and how disagreements are adjudicated.

Why abstention and dynamic testing deserve separate endpoints

A device can produce a plausible answer while failing the safer behavior, such as refusing an unsupported request or escalating an urgent one. A 2026 review concluded that current large language models still struggle to refuse inappropriate prompts (www.nature.com). Therefore, measure appropriate answer, appropriate abstention, appropriate escalation, and inappropriate silence separately.

Static correctness is also fragile. A physician-created benchmark used **2,069 open-ended items developed by 32 specialists** (www.nature.com), while dynamic robustness testing elsewhere caused **94%** of previously correct responses to fail under adaptive challenges. Agent-SafetyBench similarly reported that none of **16 tool-using agents** exceeded a **60% safety score** (arxiv.org). These studies do not set device acceptance thresholds. They show why a locked ordinary-case score cannot be the entire safety case.

Lifecycle Monitoring and Supplier Change Control

Monitoring architecture

Monitoring should join technical signals to quality-system processes. A metric without device, model, configuration, user, and outcome context cannot support investigation. NIST recommends testing before deployment and regularly during operation (airc.nist.gov); IMDRF describes clinical evaluation as iterative and continuous (www.imdrf.org).

The minimum signal chain is:

```
Prompt and clinical input
          |
          v
Configured device: model + system prompt + retrieval + tools
```

```
          |
          v
Safety layer: scope, authorization, uncertainty, escalation
          |
          +------------------> Human review or stop
          |
          v
Output or bounded action
          |
          v
Audit log: versions, inputs, tool calls, output, override, latency
          |
          v
Postmarket signal store <---- complaints, incidents, adjudication, outcomes
          |
          v
Trend rule --> investigate --> contain --> rollback or controlled change
```

Health Canada notes that transparency helps parties detect errors or declining performance (www.canada.ca). Joint principles likewise expect deployed models to support monitoring in real-world use (www.gov.uk). Monitoring should therefore measure data and workflow shift, safety behaviors, subgroup performance where observable and lawful, user overrides, escalation completion, tool errors, latency, and complaint-linked cases.

Table 3 gives an editorial trigger and responsibility model. **RACI** denotes responsible, accountable, consulted, and informed. Trigger values must be validated for the particular device.

Signal or change	Example trigger	Immediate control	RACI lead	Required record
Safety-critical miss trend	Lower confidence bound crosses prespecified limit (www.canada.ca)	Contain affected workflow; clinical review	R: Safety; A: Medical officer	Cases, adjudication, containment decision
Scope or refusal drift	Sustained rise in unsafe answer rate (www.nature.com)	Tighten routing or disable affected intent	R: Product; A: Quality	Trend chart, root cause, regression result
Population or input shift	Distance or prevalence metric exceeds validated band (www.who.int)	Stratified review and targeted retest	R: Data science; A: Clinical	Shift analysis and subgroup report
Foundation-model change	New snapshot, retirement notice, or silent behavior change (platform.claude.com)	Freeze rollout; run locked regression suite	R: ML; A: Change board	Supplier notice, impact score, test comparison
Tool or retrieval change	Schema, permission, corpus, or endpoint revision (genai.owasp.org)	Isolate change; revalidate affected hazards	R: Systems; A: Security	Dependency manifest and tool traces
Complaint-linked pattern	Prespecified severity-frequency rule met (airc.nist.gov)	Quality investigation and reporting assessment	R: Quality; A: Regulatory	Complaint linkage, version history, decision

The trigger table turns postmarket monitoring into an action system. Every signal needs an owner, investigation clock, decision threshold, containment option, and closure evidence. IntuitionLabs' first-party evidence model similarly emphasizes tracking quality, risk signals, reliability, and support burden before scaling (intuitionlabs.ai). That is implementation perspective, not regulatory authority.

Foundation-model supplier evidence

A sponsor should maintain a supplier evidence package for each released configuration:

- **Identity and lifecycle:** Exact model identifier, snapshot, region, availability status, retirement date, and replacement path (platform.claude.com).
- **Change notice:** Notice period, emergency-change terms, affected interfaces, and notification owners (docs.aws.amazon.com).
- **Evaluation:** Supplier system card, relevant safety evaluations, limitations, and sponsor-run comparative results (docs.cloud.google.com).
- **Data controls:** Retention, training-use terms, residency, encryption, access logging, and subprocessors.
- **Technical traceability:** Request identifiers, tool-call schema, rate limits, deterministic settings, and release notes (developers.openai.com).
- **Rollback:** Continued access to the prior version, exportability, routing controls, and a tested fallback (learn.microsoft.com).

Public supplier documentation shows why contracts and controls must be vendor-specific. Anthropic documents at least **60 days' notice** before retirement of publicly released models (platform.claude.com). Amazon Bedrock describes legacy periods of **six months or 45 days** (docs.aws.amazon.com). Those are provider policies, not FDA transition windows, and contract terms may differ.

Technical controls also differ. Google Cloud documents comparisons across models, versions, and evaluation jobs (docs.cloud.google.com). OpenAI recommends pinned versions and application-specific evaluations (developers.openai.com). Microsoft recommends keeping the old deployment reachable during migration (learn.microsoft.com). A sponsor should translate whichever controls actually exist into its own validated change and rollback procedure.

Editorial foundation-model change impact score

For triage, this report proposes **change impact = severity × exposure × detectability**, each scored 1 to 5. The resulting score ranges from **1 to 125**. This is an editorial prioritization tool, not an FDA formula or validated risk method.

- **Severity:** Worst credible consequence if the change introduces a failure.
- **Exposure:** Proportion and frequency of uses affected before containment.
- **Detectability:** 1 for reliably detected before use, 5 for unlikely to be detected before consequence.

The score should route work, not decide acceptability. A low aggregate score must not waive a mandatory control or mask a catastrophic single failure. IEC lifecycle material identifies data versioning and traceability as relevant dataset controls (webstore.iec.ch), which supports recording the evidence behind each factor and every reassessment.

Implementation Considerations and Process Changes

A 30/60/90-day readiness plan

The following sequence is editorial. It is designed around the September 19 publication date and the currently stated October 19 comment deadline, not around an FDA-mandated schedule.

Days 0 to 30: frame the device and file useful comments

- **Freeze the claim set:** List intended use, indications, users, environment, inputs, outputs, actions, and exclusions.
- **Place each function:** Document behavior and incorrect-output consequence on the two-axis organizer, with rationale and uncertainty.
- **Build hazard links:** Connect hazards to safety controls, competency domains, tests, acceptance criteria, and postmarket signals.
- **Lock the evaluated stack:** Record model, prompt, retrieval corpus, embeddings, tools, safety layer, infrastructure, and configuration.
- **Inventory evidence:** Mark each requirement as supported, partially supported, planned, or unsupported.
- **Submit focused feedback:** Answer only questions where the organization has evidence or a concrete implementation concern. FDA allows partial responses; breadth is not the objective.

Days 31 to 60: execute the highest-risk evidence backlog

- **Run locked non-clinical tests:** Include ordinary, edge, subgroup, multi-turn, adversarial, refusal, and tool-error sets.
- **Complete failure adjudication:** Use prespecified clinical and technical categories with blinded review where practical.
- **Pilot monitoring:** Confirm that production-like telemetry can be joined to model version, user workflow, complaints, and adjudication.
- **Exercise change control:** Simulate a foundation-model retirement, tool schema change, and retrieval rollback.
- **Draft clinical confirmation:** Specify population, comparator, endpoints, sample-size rationale, escalation behavior, and analysis plan.

Days 61 to 90: prepare a decision-quality FDA interaction

- **Close critical traceability gaps:** Ensure every high-risk claim and control has objective evidence.
- **Package unresolved questions:** Show alternatives considered, data already generated, and how FDA feedback would change the plan.
- **Rehearse evidence retrieval:** Retrieve any requirement, version, test, failure, complaint link, or release decision promptly.
- **Hold the readiness review:** Regulatory, quality, clinical, human factors, security, and ML validation should sign or record dissent.
- **Make the gate decision:** Proceed to Pre-Submission, perform targeted remediation, narrow intended use, or defer interaction.

Readiness questions for a Pre-Submission

FDA describes the Pre-Submission program as voluntary (www.fda.gov). A useful interaction asks questions specific enough for an actionable answer:

- **Risk framing:** Does the proposed placement of each function adequately reflect directiveness, supervision, and wrong-output consequence?
- **Benchmark sufficiency:** Are the task inventory, hazard coverage, test independence, subgroup plan, and acceptance criteria appropriate?
- **Clinical confirmation:** Is the proposed retrospective, shadow, simulated, or prospective design adequate for the residual uncertainty?
- **Human factors:** Are critical tasks, reliance behaviors, escalation cues, and intervention timing represented?
- **Agentic controls:** Are tool permissions, oversight checkpoints, error injection, and recovery endpoints sufficient?
- **Change control:** Which foundation-model, prompt, retrieval, or tool changes belong in a PCCP, and which would require another submission?
- **Postmarket evidence:** Are telemetry, review sampling, drift triggers, complaint linkage, and rollback governance proportionate to risk?

The packet should state the sponsor's recommended answer, supporting data, and alternative. Sending a list of broad questions without evidence simply moves internal ambiguity into the meeting.

Questions worth addressing in the public comment

A high-value comment can distinguish what is measurable from what remains ambiguous:

- How should “activity” be assigned when a function alternates between information and action across a conversation?

- What evidence demonstrates that nominal professional supervision is timely and effective?
- How should sponsors validate rare but severe safety-critical behaviors without unstable pass rates?
- When can retrospective or shadow evidence replace prospective investigation?
- How should subgroup consistency be assessed when high-risk strata are small?
- What minimum supplier information would make a voluntary foundation-model master file useful to downstream sponsors?
- Which changes to prompts, retrieval, tools, guardrails, or foundation models are suitable for a PCCP?
- What monitoring evidence should be available at submission versus generated after authorization?

Data Analysis and Evidence

The available data do not establish GenAI-specific regulatory pass marks. They do quantify why evidence architecture matters. FDA's live AI-enabled device list is not comprehensive, so counts should be treated as a changing administrative view, not market size. A 2026 peer-reviewed census systematically reviewed **1,357 devices through December 5, 2025** (journals.plos.org). **Radiology represented 78%, or 1,059 of 1,357**, limiting generalization to conversational and agentic functions (journals.plos.org).

Public clinical evidence was sparse in that census. Only **2.5%** had registered trials (journals.plos.org, **0.9%** had results posted (journals.plos.org, and **0.2%** reported patient outcomes (journals.plos.org). These numbers do not mean the devices lacked evidence submitted to FDA, because public trial linkage and regulatory evidence are different datasets. They do show that external reviewers cannot infer a mature GenAI evaluation template from public records alone.

A **691-device** study illustrates recurring reporting gaps. The median reported validation sample was **300**, with an interquartile range of **135 to 700** (jamanetwork.com), and **88.6%** did not make testing or validation data publicly available online (jamanetwork.com). These public-reporting measures are not a substitute for reviewing evidence submitted to FDA, but they show why a reusable evidence architecture must preserve sample, subgroup, version, and outcome details.

The broader clinical-AI literature also cautions against equating retrospective technical comparison with clinical benefit. An **81-study** review found only **nine prospective studies** and **six tested in real clinical settings** (www.bmj.com). Full datasets and code were unavailable in **95% and 93%** of studies, respectively (www.bmj.com). A later implementation review covered **82,656 patients and 915 clinicians** (bmjopen.bmj.com), but only seven studies reported gender and four reported PROGRESS-PLUS equity criteria (bmjopen.bmj.com).

Postmarket methods remain immature. A 2024 review included **39 sources** (pubmed.ncbi.nlm.nih.gov), with opinion or narrative reviews and simulation studies each representing **33%** (pubmed.ncbi.nlm.nih.gov. It identified only **one monitoring guideline** (pubmed.ncbi.nlm.nih.gov). That finding supports piloting signal capture before launch and documenting why selected metrics and thresholds are fit for the particular use.

The central analytical conclusion is narrow but consequential: published percentages describe evidence availability and study design, not the probability that any one GenAI device is safe or effective. The correct response is not a universal threshold. It is a traceable, risk-specific plan that makes test coverage, clinical relevance, uncertainty, subgroup behavior, and lifecycle controls inspectable. IMDRF's call for traceability and reproducibility supports that evidence discipline (www.imdrf.org).

Implications and Future Directions

The discussion paper moves the unit of analysis from the base model to the **configured user-facing system**. That implies that evidence must cover prompts, retrieval, tools, safety layers, interfaces, and human workflow. A supplier system card is useful, but cannot validate a downstream intended use. Likewise, a high benchmark score does not validate escalation, human comprehension, or an autonomous action chain. Regular operational testing remains necessary after deployment (airc.nist.gov).

The proposed two-axis framework could make evidence proportionality more explicit, but operational definitions will matter. “Supervised” should describe observable intervention opportunity, not merely the presence of a clinician. “Incorrect output” should include omission, delay, misleading confidence, unsafe refusal, improper tool use, and loss of context. Comments that bring concrete examples, measures, and alternative definitions will be more useful than endorsement or opposition alone.

The likely near-term direction is greater convergence between premarket evidence and postmarket observability. Traceability, version locks, complaint linkage, monitoring triggers, and rollback are not separate operational extras. They make the original evidence interpretable after the system or its environment changes. Joint regulator guidance emphasizes the performance of the combined human-AI team (www.gov.uk), which further supports connecting technical metrics to workflow outcomes.

No public source reviewed for this report supplies a universal competency threshold, GenAI telemetry schema, drift limit, or 30/60/90-day schedule. Product owners should treat those as controlled design decisions, justify them from intended use and risk, and preserve the rationale for FDA discussion. The highest-value investment before a Pre-Submission is often not another aggregate benchmark. It is closing the trace between claims, hazards, tests, versions, clinical evidence, and production signals.

Frequently Asked Questions (FAQs)

Is the FDA discussion paper guidance?

No. FDA says it is for discussion only and not draft or final guidance. It also says it does not communicate proposed or final regulatory expectations. Teams may use it as a design input and question set, but should not represent its concepts as binding policy. NIST's AI Risk Management Framework is also explicitly voluntary, illustrating why document status must be named precisely (www.nist.gov).

What is FDA's proposed GenAI medical-device risk framework?

It is a possible two-axis organizing heuristic. One axis describes the activity performed by the function, from informational behavior toward supervised or autonomous action. The other describes the consequence of relying on an incorrect output. Greater activity and greater consequence imply greater risk and, as an editorial implementation inference, more intensive evidence. The resulting controls should still evaluate the combined human-AI team (www.gov.uk).

What should a GenAI medical-device competency assessment include?

The proposed FDA concept combines non-clinical device benchmarking and clinical confirmation. A complete sponsor matrix should cover safety-critical recognition, scope adherence, clinical proficiency, robustness, subgroup performance, uncertainty and escalation, human factors, and agentic tool behavior. Acceptance criteria should be prespecified and risk-specific, with test sets and metrics documented (airc.nist.gov).

Does every GenAI medical device need a prospective clinical study?

The discussion paper says clinical confirmation might not require one in every case. The sponsor should justify the design from risk, novelty, residual uncertainty, and similarity between available evidence and actual use. Retrospective, shadow, simulated, adjudicated real-case, and prospective approaches offer different levels of realism and exposure. Early-stage live evaluation can reveal actual performance and human-factors issues (www.nature.com).

Is a PCCP the same as the proposed GenAI framework?

No. A **Predetermined Change Control Plan (PCCP)** is reviewed by FDA as part of a marketing submission for an AI-enabled device. FDA's final PCCP guidance applies to AI-enabled devices reviewed through the 510(k), De Novo, and PMA pathways. The GenAI paper discusses broader risk, competency, monitoring, foundation-model, and agentic questions. A PCCP may manage some changes, but does not replace the complete evidence case.

What evidence should come from a foundation-model supplier?

At minimum: exact model identity, lifecycle status, change notices, system or model cards, relevant evaluations, data handling, security controls, release notes, request traceability, and rollback options. The sponsor must still validate the configured device for its own intended use and manage gaps through contracts or technical controls. Pinned versions and application-specific evaluations are a practical baseline (developers.openai.com).

When is a GenAI function ready for an FDA Pre-Submission?

Readiness exists when the sponsor can show a stable intended-use statement, risk placement, locked system configuration, requirements-to-evidence trace, preregistered criteria, representative non-clinical results, a justified clinical-confirmation plan, human-factors and agentic controls, and an operational monitoring and rollback design. The voluntary interaction should ask specific questions whose answers would change the plan.

What should teams do before the comment deadline?

Recheck the docket, identify the FDA questions where the organization has direct evidence, and submit focused responses with operational definitions, data, and implementable alternatives. In parallel, freeze the device claim set, map risks and evidence, and prioritize the gaps that affect patient consequence or autonomous behavior.

Conclusion

FDA's discussion paper is not a new evidentiary mandate. It is a timely map of the questions the agency is considering for GenAI-enabled medical devices. Its strongest organizing idea is the connection between function behavior, consequence of error, competency evidence, and lifecycle monitoring.

For product, regulatory, quality, clinical, human-factors, security, and ML-validation leaders, the practical deliverable is an auditable evidence system. It should identify the exact device configuration, link claims and hazards to prespecified tests, distinguish ordinary proficiency from safety behavior, confirm performance in realistic clinical use, and retain the telemetry and version history needed to detect change.

The proposed risk matrix, competency table, monitoring RACI, change-impact score, and 30/60/90-day plan in this report are editorial tools. Their value is not that FDA has endorsed them. Their value is that they make assumptions visible, failures reviewable, and questions specific enough for a productive comment or Pre-Submission. A function is ready for FDA discussion when its team can explain not only how well it performs, but where it should stop, when it should escalate, how evidence maps to intended use, and what happens when the model or environment changes.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.fda.gov/medical-devices/digital-health-center-excellence/considerations-regulation-generative-ai-enabled-medical-devices-discussion-paper-and-request
2. https://www.fda.gov/media/194242/download
3. https://www.fda.gov/news-events/press-announcements/fda-seeks-public-feedback-inform-regulatory-approach-generative-ai-enabled-medical-devices
4. https://intuitionlabs.ai/articles/ai-samd-clinical-evidence-real-world-studies
5. https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0001597
6. https://jamanetwork.com/journals/jama-health-forum/fullarticle/2839236
7. https://www.nature.com/articles/s41746-025-02277-8
8. https://www.nature.com/articles/s44360-026-00152-8
9. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing
10. https://intuitionlabs.ai/articles/samd-definition-and-scope
11. https://www.imdrf.org/sites/default/files/2025-02/IMDRF_AIML%20WG_GMLP_N88%20Final.pdf
12. https://intuitionlabs.ai/articles/fda-ai-ml-samd-guidance-compliance

13. https://intuitionlabs.ai/
14. https://intuitionlabs.ai/articles/fda-pccp-implementation-guide-ai-ml-samd
15. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence
16. https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-170921-samd-n41-clinical-evaluation_1.pdf
17. https://www.gov.uk/government/publications/good-machine-learning-practice-for-medical-device-development-guiding-principles/good-machine-learning-practice-for-medical-device-development-guiding-principles
18. https://www.nature.com/articles/s41591-022-01772-9
19. https://genai.owasp.org/llmrisk/llm062025-excessive-agency/
20. https://arxiv.org/abs/2412.14470
21. https://www.nature.com/articles/s41746-026-02882-1
22. https://www.who.int/news/item/28-06-2021-who-issues-first-global-report-on-ai-in-health-and-six-guiding-principles-for-its-design-and-use
23. https://airc.nist.gov/airmf-resources/airmf/5-sec-core/
24. https://www.bmj.com/content/385/bmj-2023-078378
25. https://arxiv.org/abs/2505.08775
26. https://www.iso.org/obp/ui?_escaped_fragment_=iso%3Astd%3Aiso-iec-ieee%3A29148%3Aed-2%3Av1%3Aen
27. https://www.bmj.com/content/370/bmj.m3164
28. https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/transparency-machine-learning-guiding-principles.html
29. https://platform.claude.com/docs/en/about-claude/model-deprecations
30. https://docs.aws.amazon.com/bedrock/latest/userguide/model-lifecycle.html
31. https://docs.cloud.google.com/gemini-enterprise-agent-platform/machine-learning/evaluation/using-model-evaluation
32. https://developers.openai.com/api/reference/overview
33. https://learn.microsoft.com/en-us/azure/foundry/foundry-models/concepts/model-migration
34. https://webstore.iec.ch/en/publication/100814
35. https://intuitionlabs.ai/articles/changing-ai-models-validated-workflows-regression-testing
36. https://www.fda.gov/medical-devices/digital-health-center-excellence/resources-how-engage-fda-regarding-my-software-functions-learn-more-and-receive-feedback
37. https://intuitionlabs.ai/articles/fda-approved-ai-medical-devices-list
38. https://www.bmj.com/content/368/bmj.m689
39. https://bmjopen.bmj.com/content/13/2/e065845
40. https://pubmed.ncbi.nlm.nih.gov/39658865/
41. https://www.nist.gov/itl/ai-risk-management-framework

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai