Executive Summary of ISPE Comments

The International Society for Pharmacoepidemiology (ISPE) respectfully submits the following comments in response to the discussion paper, "Considerations for the Regulation of Generative AI-Enabled Medical Devices: Discussion Paper and Request for Feedback," issued by the Center for Devices and Radiological Health (CDRH) in August 2026 (Docket No. FDA-2026-N-7874). ISPE is a global scientific society dedicated to advancing the science and practice of pharmacoepidemiology and the use of real-world evidence (RWE) to inform decisions about human health and the use of medical products. These comments draw on the membership's expertise in observational studies, comparator selection, measurement validity, subgroup identification and assessment, the use of real-world data, and postmarket surveillance, the methods on which many of the approaches described in this paper ultimately depend. ISPE commends CDRH for seeking early input before settling on a regulatory approach, and for recognizing that evaluation methods developed for software with bounded inputs and fixed outputs may not be adequate for generative AI-enabled devices. We offer the following recommendations to strengthen the methodological rigor, clarity, and practical applicability of the approaches described.

- Apply risk assessment across the lifecycle, using worked examples and postmarket evidence to reassess risk and safeguards.
- Strengthen clinical evaluation through meaningful performance measures, validated scoring, appropriate comparators, subgroup assessment, and detailed characterization of the intended use population under routine clinical care.
- Consider rigorous non-U.S. real-world evidence as primary and/or supplementary evidence of clinical performance and safety when data are relevant and reliable and findings apply to intended U.S. use.
- Ensure actionable, feasible postmarket monitoring, with timely detection and mitigation, adequate resources, and clear manufacturer accountability.
- Use synthetic data for targeted supplementary evaluation, with real-patient validation and explicit acknowledgment of remaining evidence gaps.
- Match clinical evidence to intended use and risk, retaining adequate premarket confirmation where errors could cause serious or irreversible harm.

The following table presents ISPE's detailed comments on the discussion questions and additional points of feedback found throughout the discussion paper:

Discussion Questions	ISPE Response
Does the two-axis risk framework, organized around device activity and the consequence of relying on an incorrect output, appropriately capture the dimensions most relevant to the risk of a GenAI-enabled software function? If there are additional dimensions—such as the reversibility of a resulting action, the availability of downstream safeguards, the time pressure of the deployment setting, or the traceability of the output (i.e., to primary source materials)—that should be represented in a risk framework, please describe and provide examples of how they should be represented.	The two-axis framework provides a useful starting point for organizing the risks of GenAI-enabled device software functions. The examples accompanying the framework are helpful. We recommend extending Figure 1 into a matrix/table with worked examples across the activity categories and consequence levels. We also recommend that CDRH clarify how the framework aligns with the existing risk-based device classification system (Class I, II, and III), including worked examples illustrating how a function's position within the framework relates to device classification and the associated evidence requirements and regulatory controls. Additionally, we recommend explicitly describing the two-axis framework as a lifecycle assessment that is revisited as postmarket evidence accumulates. Initial placement on the axis would reflect the intended use and evidence available before deployment. Postmarket monitoring should inform reassessment. CDRH should clarify when such findings would warrant a change in the software function's position within the framework and when they would instead suggest additional recommendations/safeguards to support keeping its position.
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	To select and justify a clinical confirmation approach tailored to a device's intended use and proportionate to its risk profile, a sponsor would need to consider the following: - Whether a non-GenAI version of the device exists and has been previously cleared or approved for its intended use by CDRH; this would inform whether existing performance data from routine clinical care might be available to serve as a retrospective comparison group - Whether the GenAI-enabled device integrates into an existing clinical workflow or whether it introduces a new step or sufficiently changes an existing clinical workflow; this would inform whether an approach such as a shadow deployment could be used for clinical confirmation

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?	- Whether the indicated use population is well understood in terms of clinical and demographic characteristics, current standard of care, and natural history of disease; this would inform whether an approach using real world patient data could be used and evaluated for the representativeness of a selected real world data source and the corresponding patient cohort. It could also inform assessment of the representativeness of an approach such as standardized patient interactions. Finally, this could also inform an understanding of the degree of heterogeneity in clinical care for the intended use population and thus, whether clinical adjudication of real use cases is an appropriate approach.
	- Whether the GenAI-enabled device is functioning to estimate prognosis or functioning to recommend or operationally trigger clinical action.
	  - For a prognostic function that reports risk or expected course without recommending an action, confirmation should follow the prediction-model practice, that independent external validation using data that reproduce the information available at the intended prediction time, with assessment of calibration, discrimination, prediction error, and clinically important subgroups. [1]
	  - A function should enter the treatment-decision pathway when it recommends treatment, referral, timing, or dosage, or directly triggers a high-risk action. For such functions, retrospective real cases and independent specialist adjudication against the applicable standard of care can establish an output-quality and premarket safety floor, but cannot establish clinical benefit. If the device can change treatment or referral, act autonomously, or cause serious or irreversible harm when wrong, confirmation should include a prospective comparison under actual-use conditions and, when ethical and feasible, a randomized trial. Shadow deployment, standardized patients, and offline expert review may provide supporting evidence, but should not replace this evaluation.

	The approach for clinical confirmation taken by a sponsor should not only be proportionate to the device's known risk profile, according to the intended use of the device or predicate devices which are not GenAI-enabled, but should also be proportionate to the device class (tied to risk) as well as how disruptive the GenAI-enabled device is to existing clinical workflows. Sponsors should prespecify the anticipated distribution of patients, settings, users, input formats, and rare high-consequence scenarios, and should show that the confirmation data adequately cover that distribution. The labeled intended use should be limited to populations and input conditions directly supported by the evidence. Unevaluated populations or scenarios should be excluded from the initial indication and considered later through postmarket evidence and formal indication expansion rather than inferred from average performance. Where clinician oversight is claimed as a safeguard, standalone device performance and clinician-AI team performance should be evaluated separately, and the intended users' ability to detect and correct unsafe outputs should be demonstrated [2]. Postmarket monitoring may detect drift and support later expansion, but should not replace adequate premarket confirmation for high-risk treatment-decision functions. Additional methods of clinical confirmation may include the use of a real-world external control arm compared to a single-arm prospective clinical study of the GenAI-enabled device, the use of a real-world external control arm to serve as representative contextual information on the intended use population in addition to clinical adjudication of real use cases, standardized patient interactions, and/or shadow deployment. **References**

	1. Riley et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ.* 2024;384:e074820. 2.Vasey et al. DECIDE-AI. *Nature Medicine.* 2022;28:924–933.
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?	Statistically meaningful performance measurement should use prespecified measures and acceptance criteria aligned with the device's intended use and clinical risks, supported by confidence intervals rather than a single aggregate score or p-value. Diagnostic functions should follow established device and IVD principles: sensitivity and specificity against an appropriate reference standard, or positive and negative percent agreement against a non-reference comparator. Other functions should use appropriate measures, including calibration and discrimination for prognostic outputs [1], and clinically meaningful error rates and severity for open-ended outputs. Sponsors should prespecify the population, device version, sampling design, unit of analysis, denominators, and handling of unevaluable outputs and repeated observations across cases, users, and raters. Sample sizes should support adequate precision and assessment against acceptance criteria. Where applicable, standalone device and clinician–AI team performance should be reported separately. The primary performance estimate should be based on real patient data sampled or weighted to a prespecified anticipated distribution of settings, users, case complexity, missing information, and input types as aligned with the distribution of those characteristics among a real-world intended use population. Adequate precision is required for the overall intended-use population and for any subgroup-specific claim. Other subgroup analyses may be descriptive, but an anomalous point estimate should prompt investigation, additional data, or restriction of labeled use rather than dismissal as underpowered noise. Rare but high-consequence scenarios may be enriched for stress testing but should be reported separately.

	Synthetic inputs should be generated and curated independently of the evaluated device and, where a generative model is used, should come from an independent third-party model that is neither the tested device nor a closely related model. They should be used only for separately reported stress tests, should not enter overall or subgroup performance estimates, and should not support claims without real-patient anchors. The evidence needed should reflect the device's intended use, risk, and applicable regulatory pathway, including use of justified objective performance criteria or performance goals where appropriate. When benchmarking and clinical confirmation evidence are both available, they should generally be reported separately. A combined performance estimate should be considered only when the sources assess the same performance measure in sufficiently comparable populations and conditions, with methods that account for differences between sources and any overlapping cases. **References** 1. Riley et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ.* 2024;384:e074820.
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	Synthetic data may support targeted stress testing when clinically important scenarios relevant to the device's intended use are insufficiently represented in real-patient data, such as rare presentations or missing, conflicting, or atypical inputs. Testing should address identified evidence gaps and be proportionate to device risk. They may be less suitable for evaluating complex longitudinal trajectories, treatment-response relationships, rare serious adverse events, or behaviors and healthcare interactions that are poorly captured in source data.

the evaluation is intended to detect, particularly for underrepresented subgroups?	Sponsors should prespecify the purpose and estimand, disclose data provenance and generation methods, and demonstrate relevance and reliability against real-world data. Validation should assess clinically important relationships rather than only the plausibility of individual cases or similarity of marginal distributions [1]. When AI generates evaluation data, the generator should, where feasible, be independent of the evaluated device and sufficiently distinct from its underlying model to reduce correlated failures. Evaluation should include real-patient anchor cases, particularly for underrepresented subgroups. Synthetic data may help identify vulnerabilities but should not substitute for real-patient evidence supporting subgroup safety or performance claims. Limited representation should prompt additional real-world data collection and postmarket surveillance. **References** 1. Yan C, Yan Y, Wan Z, Zhang Z, Omberg L, Guinney J, Mooney SD, Malin BA. A multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications.* 2022;13(1):7609. doi:10.1038/s41467-022-35295-1.
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	ISPE recommends that CDRH align performance comparators and acceptance criteria with the device's intended use, clinical risks, and conditions of use. Comparators may include a panel of qualified clinicians or an appropriate reference standard; established objective performance criteria (OPCs) or justified performance goals (PGs) may provide acceptance targets where applicable. For open-ended outputs, sponsors should prespecify how clinical correctness, completeness, and potentially harmful errors will be assessed.

performance standard? When should human-AI team performance, rather than the device operating alone, serve as the basis for evaluation?	When qualified clinicians serve as the comparator, sponsors should define and justify whether the applicable standard represents the standard of care or the performance of a median clinician in practice. Generalist or specialist physicians should be selected according to the clinical task and expertise required. Human–AI team performance should serve as the basis for evaluation when the intended use depends on clinician interaction with the device, with standalone device performance assessed as appropriate to characterize errors and reliance on clinician oversight. Fair comparison requires holding case-mix constant (paired designs, or standardization as in provider profiling), with attention to overlap (positivity) and a specified device version (consistency), given GenAI stochasticity. Context-specific rubrics should be developed and validated prior to deployment. CDRH may suggest mixed-methods approaches incorporating both quantitative and qualitative techniques. [1] References: 1. Hong C, Chowdhury A, Sorrentino AD, et al. Application of unified health large language model evaluation framework to In-Basket message replies: bridging qualitative and quantitative assessments. Journal of the American Medical Informatics Association. 2025;32(4):626-637. doi:10.1093/jamia/ocaf023.
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	There are several approaches to assess GenAI-enabled device performance relative to the care, technology, and course of action likely to occur in the absence of the device, rather than the comparators described: - Compare to an exchangeable real-world patient population in terms of the impact on the subsequent patient journey (e.g. medications and procedures the patient receives as a result of the GenAI-enabled device)

identify and justify a comparator such as unaided clinical judgment, delayed specialist review, or no intervention?	and separately the impact on clinical and safety endpoints relevant to the intended use population. When considering a comparator such as unaided clinical judgement, delayed specialist review, or no intervention, it will be important to concretely specify the counterfactual for each approach. A well specified counterfactual (e.g. the patient outcomes if the patient is referred to a specialist 3 months later than they would having used the GenAI-enabled device) is critical to a valid and specific design of a non-active comparator of interest, of which each of these approaches would classify. Understanding the clinical and demographic characteristics, as well as standard of care treatment patterns, of these potential counterfactual populations is important when considering the degree of exchangeability between the population receiving the GenAI-enabled device (as an intervention or under routine clinical care), particularly when outside the realm of a randomized controlled clinical trial.
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?	Greater reliance on postmarket monitoring may be appropriate when residual premarket uncertainty can be reliably detected, measured, and mitigated through a robust, prospective, risk-based surveillance program. Such an approach builds on a premarket performance baseline appropriate to the device's intended use, risk, and applicable regulatory requirements and device-specific guidance, against which real-world performance can be assessed over time. The monitoring program should evaluate safety, effectiveness, reliability, input-data quality, and performance across relevant patient and clinical subgroups, while considering the device, sensing or data collection components, and AI algorithm separately. Greater reliance on postmarket evidence should reflect both device activity and the consequences of an incorrect output, as well as risk-based device classification; informational status alone should not justify reduced premarket evidence. For patient-facing functions, real-time logging should be paired with timely detection and intervention, since recording an unsafe output does not

	prevent harm. Findings should trigger reassessment of the device's risk profile and safeguards. Monitoring should combine automated, continuous surveillance with appropriately targeted human expert review and should include predefined performance thresholds, triggers, and response procedures. The balance between human review and automated monitoring of devices post-deployment should reflect the device's intended use and risk, with lower-risk systems relying more on automated monitoring with periodic human audits, and higher-risk devices using a hybrid approach combining continuous automated surveillance with expert review. Regular benchmarking against established reference datasets or appropriate validated control populations can help identify changes in calibration or performance. Monitoring should also account for changes in clinical practice, patient populations, workflows, data sources, and device components. [1] Reducing premarket evidence in reliance on postmarket monitoring may be less appropriate when device errors could cause serious or irreversible harm before detection and mitigation, even where postmarket surveillance is warranted or required. It may also be inappropriate where clinically meaningful outcomes or reference standards cannot be reliably measured in routine practice. In such cases, robust premarket evidence should remain an important component of the overall benefit-risk assessment. **References** 1. U.S. Food and Drug Administration. Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations. Draft Guidance for Industry and Food and Drug Administration Staff. January 7, 2025. Section XI: Device Performance Monitoring.

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?	ISPE supports all three approaches as complementary components of a risk-based postmarket evaluation strategy but suggests CDRH to frame postmarket monitoring as epidemiologic surveillance, not periodic re-examination. This perspective is consistent with recent calls for an epidemiology of AI that systematically measures AI exposure and its population-level health effects [1,2]. Applied to GenAI-enabled devices, this supports monitoring real-world use and associated clinical outcomes across populations and over time [3]. Each approach is a measurement instrument whose own validity must be established. Periodic re-benchmarking can determine whether performance remains consistent with the established premarket baseline, particularly when changes occur in sensors, data sources, patient populations, clinical practice, or the context of use. While periodic re-benchmarking can be useful for detecting changes in model performance, its value depends heavily on the quality and relevance of the benchmark. A change in benchmark performance is most informative when the benchmark has demonstrated construct validity for the clinical risks of interest, such as using an established reference dataset or validated control populations. [4] Public benchmarks may also become less informative over time as models are increasingly exposed to them during development or as performance approaches a ceiling. Re-benchmarking may therefore be most useful as a regression check against a prespecified baseline rather than as evidence, on its own, of continued performance in clinical practice. Re-benchmarking may also be useful when the GenAI-enabled device under routine care evolves and is no longer aligned with the initial test population. Where feasible, evaluation should rely on sequestered benchmarks with procedures to assess and limit parametric and non-parametric contamination. Sample-based clinician review offers a proximal assessment of performance in realistic clinical scenarios. However, the human reference standard introduces its own sources of uncertainty. [5] Expert reviewers may disagree when

	evaluating open-ended clinical outputs, especially for complex or ambiguous cases. Monitoring programs utilizing clinician review should pre-specify how cases and reviewers are sampled, use validated evaluation criteria, calibrate reviewer responses when possible, and report reliability metrics. Performance-degradation monitoring presents additional challenges for generative AI because many open-ended outputs do not have readily available ground-truth labels. Conventional approaches to drift detection may not reliably indicate changes in clinical performance.[6] Therefore, monitoring plans should distinguish meaningful drift, including changes to the input population and measurement between inputs and outputs. Automated performance monitoring should provide continuous or frequent surveillance for changes in input distributions, output performance, error rates, adverse events, and other relevant indicators. Additional monitoring measures can include characterization of the patient population using the devices. continued Sample-based clinician review can then provide contextual assessment of flagged cases, rare events, and clinically meaningful changes that automated metrics may not capture. Monitoring should not rely solely on fixed calendar intervals. Rather, it should combine continuous operational monitoring with more in-depth evaluation when triggers occur. These triggers might include sponsor-initiated updates, changes to an underlying model or sensor, shifts in the data environment, clinical guidelines, or practice, accumulating safety concerns, data drift, subgroup-specific performance changes, or evidence that the device is being used outside its evaluated context. Additionally, sponsors must recognize changes to the users of the tools and their environments when considering measurement validity behind the monitoring program. **References**

	1. Parikh H, McCormick T, Johnson EK, et al. The epidemiology of artificial intelligence. *Nature Health.* Published online September 15, 2026. doi:10.1038/s44360-026-00188-w. 2. Embi PJ. Algorithmovigilance—advancing methods to analyze and monitor artificial intelligence–driven health care for effectiveness and equity. JAMA Network Open. 2021;4(4):e214622. doi:10.1001/jamanetworkopen.2021.4622. 3. Feng J, Xia F, Singh K, et al. Not all clinical AI monitoring systems are created equal: review and recommendations. NEJM AI. 2025;2(2). 4. Bean AM, Kearns RO, Romanou A, et al. Measuring what matters: construct validity in large language model benchmarks. Advances in Neural Information Processing Systems. NeurIPS. 2026. https://neurips.cc/virtual/2025/loc/san-diego/poster/121477 5. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. npj Digital Medicine. 2024;7(1):258. doi:10.1038/s41746-024-01258-7. 6. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. New England Journal of Medicine. 2021;385(3):283-286. doi:10.1056/NEJMc2104626.
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?	Postmarket monitoring should draw on stakeholder expertise. Clinicians and healthcare institutions can identify clinically meaningful errors, unexpected use, and workflow effects, and link device use to longitudinal real-world data to assess downstream actions and outcomes. Professional societies such as ISPE can potentially develop common definitions, outcome measures, and evaluation methods. Standards bodies can support interoperability, device-version documentation, and consistent data capture. Monitoring should integrate into routine workflows and capture which device and version was used, the context of use, and whether outputs were presented

	and acted upon. These contributions should complement manufacturer responsibilities: manufacturers should remain accountable for maintaining monitoring systems, evaluating and investigating signals, and implementing corrective or risk-mitigation actions.

Other areas of requested feedback

Other Areas of Requested Feedback	ISPE Response
Section V.C.1, Clinical Confirmation Approaches (p. 15) — real-world data and data collected outside the United States CDRH recognizes these approaches may incorporate real-world data (RWD) and/or data collected outside the United States (OUS) and seeks input on specific considerations for the use of such data in the clinical confirmation of GenAI-enabled devices.	Methodologically rigorous studies using real-world data collected outside the United States should be considered for clinical confirmation when the data are fit for purpose (e.g. both relevant and reliable) and findings are applicable to the intended U.S. population and clinical setting. Multi-database networks and outcome validation studies highlight the importance of addressing differences in data sources, coding practices, populations, and measurement validity across settings [1,2]. Transportability (the extent to which findings apply to other populations and clinical settings) must consider differences in clinical workflows, professional roles and their designated responsibilities, treatment guidelines, language, documentation practices, healthcare access and patterns of interaction with the device, because these may modify device performance even when the underlying patient populations are clinically similar. Applicability of OUS evidence should therefore consider both transportability of the patient population and transportability of the deployment context. **References**

	1. Haber R, Webster-Clark M, Pratt N, et al. Core Concepts in Pharmacoepidemiology: Multi-Database Distributed Data Networks. *Pharmacoepidemiol Drug Saf.* 2025;34(7):e70177. doi:10.1002/pds.70177. 2. Weinstein EJ, Ritchey ME, Lo Re V. Core concepts in pharmacoepidemiology: Validation of health outcomes of interest within real-world healthcare databases. *Pharmacoepidemiol Drug Saf.* 2023;32(1):1–8. doi:10.1002/pds.5537.
Section V.C.1, Clinical Confirmation Approaches (p. 16) — synthetic data as a supplement for subgroup evaluation Where subgroup evaluation would be difficult, particularly for small or hard-to-sample subgroups where real-world data is limited, CDRH is considering that synthetically generated data modeled on real patient data could potentially supplement evaluation, assuming such synthetic data is representative of the environment in which the device is intended to be deployed.	Synthetic data may supplement evaluation of small or hard-to-sample subgroups, but particular caution is needed when limited real-world data motivate their use. Their representativeness depends on the information available to generate and validate them. Sponsors should demonstrate that synthetic data preserve clinically important relationships among patient characteristics, treatments, outcomes, and healthcare context, rather than relying only on plausible individual cases or similar marginal distributions [1]. For underrepresented subgroups, synthetic data may support targeted challenge testing and identify performance vulnerabilities, but should not substitute for real-patient evidence supporting subgroup-specific safety claims. Remaining uncertainty should be acknowledged and may warrant targeted real-world data collection and enhanced postmarket monitoring. A lack of adequate real-world data remains an evidence gap that synthetic data do not necessarily resolve. **References** 1. Yan C, Yan Y, Wan Z, Zhang Z, Omberg L, Guinney J, Mooney SD, Malin BA. A multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications.* 2022;13(1):7609. doi:10.1038/s41467-022-35295-1.

Section V.C.1, Clinical Confirmation Approaches (p. 16) — endpoints and sample size CDRH acknowledges that the approaches outlined above may not be powered around traditional effectiveness endpoints, and is interested in stakeholder input on how endpoints and sample sizes might be determined.	Endpoints and sample sizes should reflect the device's intended clinical use, the purpose of clinical confirmation, and the consequences of error. Alongside technical and clinical performance, evaluation should consider downstream effects on clinical decisions, time to appropriate care, healthcare utilization, and patient outcomes. Sample-size planning should address precision around clinically consequential error rates, including uncommon serious errors that aggregate accuracy may obscure [1]. Rare, high-consequence scenarios may warrant targeted or enriched evaluation, with results reported separately from routine-use performance estimates. Clinically important subgroups and the precision needed to support intended-use claims should be prespecified. Postmarket real-world evidence may complement premarket evaluation for outcomes or populations that cannot feasibly be assessed with sufficient precision before deployment. **References** 1. Riley et al. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. *BMJ*. 2024;384:e074821.
Section VI.A, Postmarket Monitoring Approaches (p. 20) — feasibility and practicality of implementation CDRH seeks feedback on these approaches, as well as the feasibility and practicality of implementing these or other approaches to postmarket monitoring for GenAI-enabled devices.	ISPE recommends that CDRH consider implementation feasibility using an established framework such as the Nonadoption, Abandonment, Scale-up, Spread, and Sustainability (NASSS) framework [1]. NASSS helps explain how a technology, its users, healthcare organizations, and the wider health system interact to support or hinder successful implementation over time. Applying such a framework could help identify where monitoring strategies are most likely to encounter practical barriers over time. The success of monitoring depends on clinical workflow, incentives, data access, and

	organizational capacity, in addition to technical measurement strategies. ISPE therefore recommends that CDRH consider two implementation principles. First, monitoring plans should address the practical requirements for implementation, including data access, workflow integration, staffing, and heterogeneity across deployment settings, in addition to specifying performance metrics. Second, guidance should acknowledge the role of health systems, professional societies, standards organizations, and other ecosystem partners that may contribute data, expertise, or adjudication while maintaining clear manufacturer accountability. **References:** 1.Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. *J Med Internet Res.* 2017;19(11). doi:10.2196/jmir.8775.