American College of Radiology

December 1, 2025

Food and Drug Administration
Center for Devices and Radiological Health
10903 New Hampshire Ave
Silver Spring, MD 20993

RE: Docket ID: FDA-2025-N4203) Request for Public Comment: Measuring and Evaluating Artificial Intelligence-enabled Medical Device Performance in the Real-World; Comments of the American College of Radiology

The American College of Radiology (ACR)—a professional association representing more than 40,000 physicians practicing diagnostic radiology, interventional radiology, radiation oncology, and nuclear medicine, as well as medical physicists—appreciates the opportunity to comment on the Food and Drug Administration's (FDA) request for public comment, "Measuring and Evaluating Artificial Intelligence-enabled Medical Device Performance in the Real-World" (FDA-20250N4203) published September 30, 2025.

ACR Data Science Institute

ACR established its Data Science Institute (DSI) in 2017 to empower the advancement, validation, and implementation of artificial intelligence in medical imaging and the radiological sciences for the benefit of our patients, society, and the profession. The DSI collaborates with radiology professionals, industry, government, patients, and other stakeholders to develop programs and tools in support of the safe and effective implementation of AI applications to improve patient care. To date, DSI initiatives include:

- Defining clinically relevant use cases intended to guide the development of useful imaging AI (ACR Define-AI).
- Establishing and making broadly available the first national recognition program for safe and effective implementation of AI in imaging practices, which is now used in many different sites across the U.S. to guide proper process design and ongoing quality assurance of imaging AI implementation (ACR ARCH-AI).
- Creating opportunities to monitor the effectiveness of AI models in real-world clinical practice, including the first large-scale quality registry for AI performance monitoring (ACR Assess-AI).
- Founding member of the Healthcare AI Challenge consortium, a multi-institution collaborative effort dedicated to crowdsourced evaluation of generative AI solutions.
- Instituting a comprehensive, regularly updated online portal with an overview of all commercially available, FDA cleared SaMD products, including information of training and characteristics of radiology AI models (including model cards). This

helps radiologists choose what works for their practices and patients (ACR AI Central).

- Organizing thought leadership activities regarding the regulatory, legal, and ethical issues associated with radiology AI.

Responses to FDA Questions:

1. Performance Metrics and Indicators

1(a). What metrics or performance indicators do you use to measure the safety, effectiveness, and reliability of AI-enabled medical devices in real-world clinical use?

Evaluating the safety, effectiveness, and reliability of AI-enabled medical devices in real-world clinical settings requires a comprehensive, risk-informed approach that incorporates multiple performance indicators. Key metrics include critical miss rates for time-sensitive findings, such as intracranial hemorrhage or pulmonary embolism, and near-misses identified and overridden by qualified end-users. Diagnostic performance measures—including sensitivity, Receiver Operating Characteristic (ROC) Curve–Area Under the Curve (AUC), and positive and negative predictive values (PPV and NPV)—are essential for assessing accuracy. Additionally, concordance and discordance rates between algorithm output and final radiologist reports, which serve as the primary clinical reference standard, should be calculated and tracked on an ongoing basis.

Demographic subgroup error rates should be monitored to confirm that no patient population is disproportionately affected by differences in model performance. Operational metrics such as workflow outcomes, including turnaround times and recall rates, also provide insight into the device's impact on clinical efficiency. Finally, ongoing system reliability and monitoring—covering uptime and availability, time-to-inference, and version tracking—supports transparency and traceability throughout the device lifecycle. Monitoring of machine learning operations must also include systematic vigilance and capture of overall delays, unsuccessful model processing, and failed routing of data, earmarked exam codes to matching models, and complete or appropriately timed dispatch of source data from scanner to AI infrastructure.

1(b). How are these metrics defined, and weighted when assessing different dimensions of performance and safety?

Performance and safety metrics would be weighted differently based on use case and risk. For example, the tolerance for *false negatives* may be lower in some use cases, particularly for critical findings. The tolerance for *false positives,* on the other hand, may be lower when a radiologist must verify multiple AI results/inferences, which could cause delays in patient care. Any failure in proper flow, routing, and processing of data or other clinically unacceptable process delays requires systematic root cause analysis and rectification on an ongoing basis. This vigilance is particularly important as the ecosystem of scanners, protocols, and exam and component series names will constantly change over time.

1(c). What timeframe do you consider when evaluating “real-world clinical use” performance?

There should be local acceptance testing periods for each phase of integration. For example, some institutions implement a three-phase review over the course of approximately 3 to 6 months. First, model performance is assessed in a sandbox environment on a specific set of cases with known ground truth. If the model performs as expected, a limited evaluation in a clinical environment on day-to-day patients would be next. If both model performance and user experience are positive, the model can be released into full production, enabling ongoing performance monitoring in a clinical environment across all users. Model version updates over time require re-assessment and must be recorded in the monitoring database to allow review and interpretation if performance shift occurs.

2. Real-World Evaluation Methods and Infrastructure

2(a). What tools, methodologies, or processes are you currently using to proactively monitor AI-enabled medical device performance post-deployment?

ACR’s Assess-AI national registry provides qualified end-users with technology for continuous automated monitoring of model inputs and outputs, versioning, and concordance with radiologist reports, allowing users to capture data during clinical use to ensure safe and effective performance of AI medical devices. In addition, Assess-AI compares performance of local model-based inference with aggregated national data (processed by same or different commercial models on identical clinical use cases) to uncover performance variation and identify improvement opportunities. Forensics application tools provide detailed analysis and reporting in conjunction with local options for re-identification of source data. As a result, clinical sites can support their local tools, methodologies, and processes to identify and fix quality issues, including those requiring adjustments to image acquisition technique.

2(b). How do you balance human expert review and automated monitoring approaches in your evaluation methodology, and what are the pros and cons of each when it comes to practical implementation?

Automated data capture/monitoring is crucial for ongoing, at-scale evaluation of the performance of AI medical devices, both to establish ground truth and to identify trends and potential inconsistencies. Due to clinical time constraints and the high volume of transactions occurring in a busy clinical practice, monitoring of AI devices is not possible without initial automation. Automated monitoring and determinations of ground truth can be flawed, however, and human adjudication with expert review of discordant cases is necessary. Furthermore, automated (AI-based) labeling methods require careful initial assessment by human experts during the design and testing stage to ensure robust performance within the desired clinical/use case context. When the automated assessment flags performance issues, local human expert review can identify underlying issues and make decisions about informed adjustments to address the clinical needs.

2(c). What technical, operational, or organizational infrastructure supports your real-world AI-enabled medical device performance evaluation?

From a technical perspective, ACR developed TRIAD (Transfer of Images and Data), a free, cloud-based data collection and exchange toolset that has been utilized by over 20,000 radiology facilities to support clinical trials, registries, national accreditation, and education and quality programs. ACR Connect, the next generation of TRIAD, serves as a common base platform. ACR Connect is integrated with local production systems (PACS, RIS etc.) and can host specialized functional software modules for local data processing and integration with ACR programs, including the data registry application (app), clinical research app, and a federated AI network app. Clinical sites can use the ACR Connect software to participate in ACR's national AI registry, Assess-AI, to monitor AI performance in practice. ACR Connect is currently active with 2,321 facilities participating, and usage of the software is growing.

On the operational side, ACR leverages its National Radiology Data Registry (NRDR) program, which is also on the ACR Connect platform and offers mature registry operations and services to participating sites. These services include governance structures, business associate agreements, rigorous data quality control processes, and advanced analytics capabilities to ensure reliable and actionable insights.

Finally, at the organizational level, ACR's AI quality assurance program (ARCH-AI) promotes and recognizes adherence to best practices for AI governance in imaging interpretation while facilitating gap analysis and quality and process improvement for prospects. ARCH-AI ensures that sites maintain high standards for safety, accountability, and transparency when deploying AI tools in clinical workflows. ACR volunteers and staff members are currently working on a draft Practice Parameter for Imaging AI, which will form the basis for future accreditation program design.

3. Postmarket Data Sources and Quality Management

3(a). What data sources do you typically use for ongoing performance evaluation (e.g., electronic health records, device logs, patient-reported outcomes)?

Currently, core data sources for ongoing performance evaluation must include imaging AI outputs combined with device and exam metadata, which provide essential details about algorithm performance in specific scenarios. User interaction logs, such as whether qualified end-users accept or override AI (if applicable to the interaction), can offer insight into performance as well as engender trust.

Patient location (e.g., inpatient, outpatient, emergency) and population demographic data are necessary to help monitor performance across different subgroups. Additionally, final radiology reports serve as the primary clinical reference standard to validate AI outputs against expert radiologist interpretation. In some cases, additional data from the electronic health record (EHR), including outcomes reported by other clinicians on the care team, can also be considered to provide a broader view of patient impact beyond imaging findings. Some of this data is currently accessible and culled from the DICOM image header. In our approach, we collect the non-PHI portion of the metadata from the source images without collecting the actual images. In the future, we plan to establish an interface option to the local EHR capable of collecting additional clinical data; this will enhance the ability to adjudicate model performance.

3(b). How do you address data quality, completeness, and interoperability challenges in your monitoring systems?

The ACR Assess-AI registry uses a standards-first ingestion pipeline (described in our specification document) based on a shared data dictionary and common data elements. It accepts AI outputs, radiology reports, and imaging metadata via ACR Connect using HL7 v2, DICOM, and registry-defined flat files. Incoming data are validated locally through schema checks, required-field enforcement, and status gating. Only complete, final reports and conformant records are accepted. At the site level, this system provides radiologists, informatics teams, and quality leaders with a structured way to confirm all eligible patients and studies for a given use case are captured, submissions are tracked, and exceptions are reviewed. This process supports local QA, accreditation, and interoperability goals, while also contributing to the national registry.

Participating sites process and de-identify data locally using ACR Connect. PHI is kept on premises or in the site’s cloud, and only limited, standards-conformant datasets are submitted to ACR. Assess-AI tracks algorithm versions, inputs, and concordance between AI outputs and radiology reports over time and across sub-cohorts. Local clinical governance committees can use the resulting dashboards to benchmark performance against national peer organizations, investigate outliers, and detect drift early. In many practices, these reports are reviewed alongside existing quality and safety metrics. This allows the local teams responsible for clinical care to act on AI performance signals within a familiar governance framework.

3(c). What methods have been most effective in incorporating clinical outcomes and user feedback into model updates?

Sites and manufacturers can employ a multi-layered approach to ensure that clinical outcomes and user feedback inform model improvements in an effective manner. Concordance analytics can be used to compare AI inferences/outputs with final radiologist reports. When discrepancies arise, drill-down analyses help identify patterns and root causes. Additionally, issue logs from clinical sites capture real-world challenges and drive vendor corrective actions and site-level remediation plans. These insights can be compiled into structured reports to share with both developers and participating clinical sites. This process creates a continuous feedback loop that informs model refinement to improve reliability. Some vendors have data use agreements with local sites to use source images and final reports for training and improvement of future versions of the models.

4. Monitoring Triggers and Response Protocols

4(a). What triggers the need for additional assessments and more intensive evaluation?

Several factors can trigger the need for additional assessment and evaluation of AI-enabled medical devices in radiology practices. One key driver is unmet performance expectations, such as breaches of established performance thresholds, delays that exceed the maximum response time defined in the Service Level Agreement (SLA), or instances where the system does not produce any output.

Another important consideration is significant infrastructure change, such as the introduction of new clinical software or hardware that presents input data to the model or a change in the imaging protocol. These modifications can affect device performance and require retesting/revalidation.

Finally, changes in patient populations—such as shifts in referral patterns or the addition of new care sites through hospital acquisitions or clinic openings—may alter the data distribution and clinical context. These changes may necessitate further evaluation to ensure continued AI accuracy and reliability.

4(b). How do you define and respond to performance degradation in real-world settings?

Governance frameworks should clearly define what constitutes performance degradation and establishes a structured, multi-faceted response process. When degradation is detected at a site, common response actions include notifying the vendor to ensure awareness and initiate any contractual remedies. Organizations may conduct targeted shadow audits to confirm the root cause and confirm vendor compliance with regulatory and contractual obligations. In cases where patient safety or operational integrity is at risk, a temporary rollback or suspension of user access may be implemented to prevent further impact. Following resolution, a post-remediation evaluation period is essential to confirm system stability before full restoration of clinical use.

5. Human-AI Interaction and User Experience

5(a). How do clinical usage patterns and user interactions influence AI-enabled medical device performance over time based on your observations?

Clinical usage patterns and human–AI interaction significantly influence how an AI-enabled device performs after deployment, even when the underlying model remains the same. The placement of the tool within the workflow (triage, concurrent support, or post-read QA) affects accuracy, time-to-action, and documentation practices. Alert volume, batching, and overnight use impact override rates and concordance due to fatigue and automation bias. Performance also varies with scanner/protocol updates and local factors like patient mix, exam types, and equipment. These examples highlight the importance of ongoing monitoring of concordance, override rates, and subgroup performance, along with retraining and revalidation whenever usage patterns or inputs change.

5(b). What design features, user training, or communication strategies have proven most effective for maintaining safe and effective use as systems evolve?

Involvement of relevant clinical stakeholders is paramount. In the context of imaging AI this often involves identifying clinical champion(s) and subject matter experts in each practice. Informed selection of the best suited model to address a real-world clinical challenge in the practice is next. This selection is optimally based on episodic testing of the product(s) on retrospective data from the same practice, or in a limited trial run which also serves to educate local champions about the performance, user interface, and other relevant items regarding the software. Another important building block is obtaining and issuing the Instructions for Use (IFU) to end users. If the IFU is not suitable to educate end users, the local team may decide to produce additional instructional materials (with or without the assistance of the vendor).

Effective design of the software itself begins with clear user interface (UI) affordances, such as confidence indicators and clear distinctions between actionable and advisory outputs. Audit trails for accept or override decisions provide accountability and support ongoing quality improvement. Explainability features in the UI should also offer meaningful context without creating a false sense of precision. Additionally, the AI should be thoughtfully deployed within the clinical workflow.

Training and retraining are also critical (see above initial comments). Role-based programs should be tailored for clinical end-users and other personnel and tied to documented use cases and fallback procedures. These programs should emphasize review and validation of outputs with clinical/professional judgement.

Finally, transparency builds trust and mitigates misuse. Providing qualified end-users with clear information at the point of use—such as indications, training populations, and system limitations—adds to their confidence and understanding of whether a given AI tool can safely and effectively support the use case.

6. Additional Considerations and Best Practices

6(a). In addition to the factors previously mentioned, what other considerations, best practices, or tools were important in the development and implementation of your real-world validation system?

ACR's ARCH-AI, which recognizes healthcare organizations that follow best practices for clinical AI usage in medical imaging. Supportive tools, such as the ACR Assess-AI registry, use a practical lifecycle approach to AI quality assurance. This approach aligns key components of responsible AI implementation—good governance, model selection, acceptance testing, and continuous monitoring—and signals accountability to patients and payers. ACR's programs leverage NRDR's mature registry backbone and ACR Connect to provide secure, standardized, and scalable real-world evidence generation. Additionally, use of ACR's AI Central device database informs and improves AI procurement and deployment decisions. Finally, community knowledge and experience sharing between sites (e.g., ARCH-AI Learning Network) supports the development and evolution of best practices for AI evaluation, deployment, and monitoring in real-world environments.

6(b). Please address any implementation barriers encountered, incentives that supported your efforts, and approaches to maintaining patient privacy and data protections.

Barriers to effective implementation of AI-enabled devices include integration burdens, lack of common results semantics, and uneven access to EHRs and outcomes. Some sites may also have uncertainty regarding post-market expectations for monitoring and reporting, including sharing data with AI performance monitoring mechanisms. While there are IHE standards available to help address these barriers, they are not widely implemented across practices.

Additional challenges arise when the AI solution performs a quantitative analysis which cannot be validated simply by the interpreting radiologist or would otherwise require significant manual effort that is not feasible in daily practice. In addition, quantitative outputs such as risk scores may only be confirmed over time if clinical cases are tracked. Both scenarios represent challenges for performance monitoring approaches, as they cannot be easily solved with the currently available infrastructure and approach. Longitudinal tracking of software performance over time on the same patient use case can provide insights into performance and aid future monitoring programs.

FDA should recognize participation in qualified AI registries—for example, ACR's Assess-AI on NRDR infrastructure—as helping to fulfill manufacturers' post-market surveillance commitments. Incentivized participation can identify drift in model performance across various practice settings and promote readily available tools for continuous monitoring.

Conclusion

The ACR welcomes further communication and collaboration with the FDA on AI device safety and effectiveness. For questions, please do not hesitate to contact Michael Peters, ACR Senior Director, Government Affairs, at mpeters@acr.org or Lindsay Robbins, ACR Regulatory Policy Specialist, at lmrobbins@acr.org.

Sincerely,

D Smetherman

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