RADIOLOGICAL SOCIETY OF NORTH AMERICA
820 JORIE BLVD., SUITE 200, OAK BROOK, IL 60523-2251
TEL 1-630-571-2670 FAX 1-630-571-7837
RSNA.ORG

December 1, 2025

Re: Docket No. FDA-2025-N-4203, Measuring and Evaluating Artificial Intelligence-enabled Medical Device Performance in the Real World; Request for Public Comment

The Radiological Society of North America (RSNA) is a leading global organization representing over 52,000 radiologists and medical imaging professionals across 150 countries. Radiology and medical imaging are among the most data-intensive fields in medicine, and AI-driven technologies have already begun transforming clinical practice. Radiology has experienced the highest rate of medical AI tool development and deployment, with more than 76% of the over 1000 FDA-cleared AI algorithms designed for radiological applications.

RSNA appreciates the opportunity to provide the following comments in response to FDA's request for input on Docket No. FDA-2025-N-4203, Measuring and Evaluating Artificial Intelligence-enabled Medical Device Performance in the Real World.

Performance Metrics and Indicators

1a. 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?

RSNA has curated large, expertly annotated datasets (https://mira.rsna.org) across modalities (e.g., chest radiograph, CT, MSK, abdominal imaging) that include training, validation and sequestered test partitions. These datasets provide benchmarks for public evaluation and validation of AI-enable medical device effectiveness and reliability. Models are scored based on metrics such as log loss, area under receiver operating characteristic curve, and sensitivity/specificity to assess performance and generalizability.

Together, these metrics and benchmarks form RSNA's framework for continuous performance monitoring, transparency, and safety assurance, aligned with FDA and NIH priorities for lifecycle evaluation of AI-enabled medical devices.

Metrics and performance indicators for AI-enabled medical devices in radiology include:

- Accuracy (sensitivity, specificity, area under the curve [AUC])
- Bias/fairness across demographic groups
- Workflow impact
    - Technical efficiency of algorithm execution
    - Usability of AI results for end user

- Stability over time as evidenced by indicators such as:
  - Calibration drift
  - Error rates
  - Diagnostic concordance
  - User trust metrics
- Reliability as assessed through consistency across imaging devices and manufacturers, modalities, protocols and acquisition conditions, supported by structured feedback from clinicians.

Future AI benchmarks/assessments should incorporate multidimensional evaluation such as:

- Diagnostic accuracy
- Execution boundary conditions and failure modes
- Explainability and chain-of-reasoning quality
- Communication clarity, and consistency

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

AI performance metrics are defined through a multidimensional framework emphasizing clinical validity, technical robustness and operational reliability and consistency. Operational reliability is based on high service availability and minimal downtime. Consistency involves producing the same results in response to the same inputs over time. Continuous, post-deployment auditing with benchmarks, coupled with human-AI comparative review.

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

Real-world performance is tracked monthly through benchmarks that compare local model performance against performance of the same model at peer institutions. Human reviews are performed annually to detect drift. Reassessments are also regularly performed following software upgrades, model or workflow updates, scanner protocol changes, or shifts in clinical practice that can affect model performance.

Real-World Evaluation Methods and Infrastructure

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

Continuous monitoring dashboards populated by statistical tools that measure performance and detect drift are used to monitor AI-enabled medical devices after deployment. Dashboards provide trendlines on AI performance and can be filtered by device, location, patient population and other parameters.

2b. 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?

Expert review and automated monitoring approaches complement one another. Automation using a generative model or ensemble of models enables scalability and early alerts. Regular expert review ensures context and more granular detail.

Automated and expert-driven evaluation are balanced through a layered validation framework. Automated methods provide continuous tracking of quantitative indicators, such as accuracy drift, calibration stability, and dataset representativeness. Human expert review complements this automation by ensuring clinical relevance and contextual accuracy, and it also offers an escalation pathway for automated monitoring in when problematic or discrepant cases have been flagged. Radiologists and other domain experts evaluate outputs for diagnostic plausibility, fairness, and usability, factors that automated metrics currently cannot adequately capture.

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

Data analytical tools are used to monitor performance and detect statistical drift compared to industry benchmarks and to populate continuous monitoring dashboards. Model outputs are stored in data structures that can be analyzed with data tools and compared to benchmarks and expert-determined ground truth.

Postmarket Data Sources and Quality Management

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

Standardized annotated datasets with multi-expert consensus serve as ground truth for AI performance evaluation. AI benchmarks are based on evaluation of model outputs against human-curated ground truth datasets incorporating radiologist clinical observations and relevant clinical variables including pathology and laboratory results.

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

Radiology sites address interoperability challenges by requiring that systems comply with established standards (DICOM, HL7 FHIR, IHE) and standardized APIs. Data quality and completeness are improved by the use of consistent terminology for procedures, protocols, reporting and AI results (RadLex, LOINC). Data quality and interoperability must be evaluated and enhanced through periodic human review.

Monitoring Triggers and Response Protocols

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

The need for additional assessment of AI performance is monitored with statistical process control (SPC) rather than single-point fluctuations. Immediate triggers can be defined as observation of variance by multiple standard deviations from the historical mean, or as sustained shifts indicated by multiple consecutive points on one side of the mean. Response protocols should include baseline recalculation as well as efforts to redress after a change in environment such as scanner upgrade or changes in scope of implementation.

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

In clinical settings, performance degradation is addressed using root cause analysis, such as reviewing non-AI model-related causes first. It is very important to differentiate between degradations in actual AI inference due to drifts in data or concept and degradations from far more common challenges such as new scanners with unrecognized metadata reducing the effectiveness of preprocessing systems, or new users entering the health system (e.g. per diem or locum) that use the AI solution inappropriately.

When the root cause is determined to be the AI model, the focus is to identify what has changed since the model's implementation: the input data (data drift), or the ground truth (concept drift). Responding to performance degradation may require disabling a model and reintroducing after confirming that the performance issue has been corrected.

Human-AI Interaction and User Experience

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

In radiology, clinical usage patterns and user interactions can significantly shape AI performance over time. It is imperative to measure user engagement with the AI tool and outputs in order to gain a full understanding of user interaction and influence. Overreliance or automation bias can degrade the quality of expert oversight. Variable image acquisition practices and differing EHR integration pathways can alter input data characteristics, which affects the consistency of the overall human-AI diagnostic system. Monitoring override rates, reading times, and histopathology correlates can help identify such patterns.

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

In radiology, measures for maintaining safe and effective use of AI systems over time include interface and workflow design that embeds the AI tool in the radiologist workflow, while separating AI-generated annotations from primary clinical images within the viewing system. This prevents unverified AI overlays, especially false positives or negatives, from influencing less-trained users. Displaying uncertainty indicators and offering concise explanations also fosters trust and safe use. Clear feedback channels sustain effective adoption while maintaining trust.

Additional Considerations and Best Practices

6a. 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?

RSNA has collaborated with other radiology professional associations to publish a set of principles for the use of AI emphasizing transparency, clinical validation, and reproducibility ("Developing, Purchasing, Implementing and Monitoring AI Tools in Radiology: Practical Considerations. A Multi-Society Statement From the ACR, CAR, ESR, RANZCR & RSNA," Brady et al [https://pubmed.ncbi.nlm.nih.gov/38276923/]). Through its work in producing training and validation datasets and establishing performance benchmarks, RSNA has made available a set of quality assurance tools for consistent assessment of AI performance. Governance frameworks should mirror the American College of Radiology (ACR) accreditation and quality-control structures, ensuring traceability and peer oversight.

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

Barriers to implementation include lack of AI transparency, fragmented data, variability in labeling, and uneven access to real-world validation.

RSNA values the opportunity to provide these comments and looks forward to continued collaboration with FDA. For additional information or questions, please contact RSNA's Director of Government Relations, Libby O'Hare (eohare@rsna.org).

Sincerely,

Jeffrey Klein

Jeffrey Klein, MD
Chair of the Board
Radiological Society of North America