Response to:

US Food & Drug Administration Docket No. Docket FDA 2025-N-4203

Request for Public Comment: Measuring and Evaluating AI-enabled Medical Device Performance in the Real-World

Submitted by the
International Society of Pharmacoepidemiology

December 1, 2025

Request for Public Comment: “Measuring and Evaluating AI-enabled Medical Device Performance in the Real-World”

The International Society for Pharmacoepidemiology (ISPE) is pleased to have the opportunity to offer its perspective and suggestions, and submits for your consideration the following response to the FDA Request for Public Comment: “Measuring and Evaluating AI-enabled Medical Device Performance in the Real-World”.

ISPE is an international organization dedicated to advancing the health of the public by providing a global forum for the open exchange of scientific information and for the development of policy, education, and advocacy for the field of pharmacoepidemiology, including such areas as pharmacovigilance, drug utilization research, comparative effectiveness research, and therapeutic risk management. ISPE is committed to providing an unbiased scientific forum to the views of all parties with interests in drug development, drug delivery, drug use, drug costs, and drug effects.

ISPE members represent the various scientific disciplines involved in studying drugs. Members are employed by the pharmaceutical industry, academic institutions, government agencies, non-profit and for-profit private organizations. Members have degrees in a number of fields, including epidemiology, biostatistics, medicine, nursing, pharmacology, pharmacy, law, health economics, and journalism. With members in 53 countries and national chapters in Argentina, Belgium, Denmark, and the Netherlands, ISPE truly provides an international forum for sharing knowledge and scientific approaches to foster the science of pharmacoepidemiology. We thank the FDA for allowing us the opportunity to comment on this document. ISPE welcomes any future dialogue with the FDA.

Sincerely,

International Society for Pharmacoepidemiology (ISPE)

Question 1. Performance Metrics and Indicators

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?*

These metrics are distinct for three categories of consideration: 1) the devices themselves, 2) sensing or data collection portions of the device, and 3) the AI algorithm embedded in the devices. Each of these categories merit separate consideration.

The sensing elements themselves need to have surveillance for appropriate signal collection, as deviations in the sensing data itself disrupts the AI components. An example of this is lead impedance in pacemakers/defibrillators, as the AI in these relies on accurate sensing data to perform its functions. Other key examples are the bias estimates from the pulse oximeters that the FDA has issued specific guidance on, which had lower performance in patients with darker skin. In that case, the metrics involved accurate analysis of light signals on the skin.

For the AI components, conducting analysis of the algorithm and surveillance is important not only during development but after product release.[1]

There are several frameworks and considerations for the conduct of AI surveillance, including FAIR-AI[2] and those referenced directly from the TRIPOD+AI[3] and TRIPOD-LLM[4] documents.

Similar to all medical devices, the following indicators are recommended for safety evaluation:

- Adverse event rates and the types of events
- Intrinsic device effects
- Learning curve effects
- Patient outcomes
- Operator feedback and tracking of usability issues
- Regulatory compliance metrics, if any

---

[1] Andersen ES, Birk-Korch JB, Hansen RS, Fly LH, Röttger R, Arcani DMC, Brasen CL, Brandslund I, Madsen JS (2024). Monitoring performance of clinical artificial intelligence in health care: a scoping review. *JBI evidence synthesis*, *22*(12), 2423–2446. https://doi.org/10.11124/JBIES-24-00042

[2] Wells BJ, Nguyen HM, McWilliams A, *et al.* (2025). A practical framework for appropriate implementation and review of artificial intelligence (FAIR-AI) in healthcare. *npj Digital Medicine,* 8, 514. https://doi.org/10.1038/s41746-025-01900-y

[3] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, *et al.* (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ,* 385, e078378. https://doi.org/10.1136/bmj-2023-078378

[4] Gallifant J, Afshar M, Ameen S, *et al.* (2025). The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine,* 31, 60–69. https://doi.org/10.1038/s41591-024-03425-5

Both clinical and cost effectiveness are evaluated with the following measurements:

- Clinical outcomes
- Treatment effects in the treated and control groups
- Cost effectiveness
- Adoption and adherence rates

Depending on the nature of the medical device, the following metrics are recommended to evaluate reliability:

- Uptime and availability rates
- Failure rates and the type of failure
- Mean time between failures (MTBF)
- Mean time to repair (MTTR)
- Error rates and error types
- Regular maintenance and update metrics
- Recall rate
- Explanation rate for implantable devices
- Real-world clinical performance metrics for AI-enabled diagnostics (e.g., measures of accuracy, precision, and recall)

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

The metrics are typically defined through a combination of regulatory guidelines, industry standards, and clinical expertise. For AI-enabled medical devices, metrics might be defined as adverse event rates (e.g., number of events per 1000 patient-days or per procedure) or uptime (e.g., percentage of time the device is operational over a given period). ROC-AUC and standard metrics such as precision, accuracy, and sensitivity or recall are often assessed to inform safety threshold setting.

A weighted scoring system can be used, in which metrics are assigned scores based on their relative importance. Weighting of metrics varies depending on the specific use case and priorities. From a clinical perspective, effectiveness metrics like clinical outcomes are prioritized. Weighting importance of performance and safety must be done in the context of use. For example, in pulse oximetry, 88% is the key threshold for the need for supplemental oxygen. Therefore, AI that assesses oxygenation through assessment of light reflection of skin needs to be more accurate around the threshold effect. Safety is also relative to the core context of use and the severity of adverse events relative to the monitoring. Careful monitoring for failure modes should be weighted to the level of expected patient harm if they occur.

c. *What timeframe do you consider when evaluating "real-world clinical use"*

performance?

Real-world clinical use generally refers to the period after the device is approved by the FDA and on the market ongoing routine clinical use. An argument could be made for pre-market trial data to be included, as real-world data is generated in many of these trials, but the context of data collection, use, the controlled environment and patient selection, argue for a clear separation of these data. Real-world data continue to be collected on patients throughout the remainder of the patient's use of the device, with no time limit, and censored by death or a context-relevant window following removal of exposure to the device.

The timeframe for evaluating real-world clinical use varies depending on the device, its intended use, and regulatory requirements. However, evaluation should be an iterative process, with both time-based (e.g., quarterly or annually) and trigger-based (e.g., major practice changes, software updates) reassessments. While local-level monitoring can capture site-specific behavioral shifts, aggregated, systemic evaluations can detect broader trends in drift. Continuous benchmarking using known reference datasets or validated control populations helps ensure calibration and comparability over time.

Question 2. Real-World Evaluation Methods and Infrastructure

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

In the practice of pharmacoepidemiology, comparative effectiveness analyses are usually utilized to compare the treated group (i.e., care involving the medical device) to the control group (i.e., care involving a comparator device or treatment modality is applied) for a treatment effect.. Patient, clinician and, operator factors are often controlled for to assess the intrinsic device effect. Techniques such as propensity score matching, inverse propensity treatment weighting (IPTW), weighting by the odds, sequential probability ratio testing (SPRT) and risk-adjusted survival methods are used. Active surveillance tools can also be utilized to monitor safety of a marketed medical device, such as the DELTA (Data Extraction and Longitudinal Trend Analysis) system, which utilizes accruing clinical data repositories from electronic health record systems and clinical registries.[5] DELTA is an open-source system which has been validated to support near real-time active medical device safety surveillance across diverse data environments.

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?*

[5] https://www.mdepinet.net/delta

The balance between human and automated review should align with the device's intended use and risk classification. Postmarket strategies benefit from both approaches.

Automated surveillance of performance can be done systematically and continuously across settings, over time, and should be utilized wherever possible and reasonable for the context of use. For low-risk systems, automated drift detection with periodic human audits may suffice. Automated monitoring is fast and consistent but may miss contextual factors and be sensitive to shifts in data. In many cases, the data may have biases, missingness, or gaps that create false positive safety signals that require further evaluation or generate more root cause analysis.

For moderate- to high-risk devices, hybrid approaches, where automated systems flag anomalies for domain experts to adjudicate help maintain safety and transparency. Human expert review offers contextual understanding, detects nuanced issues, and captures rare events, but is time-consuming, resource-intensive, and subject to human bias. Oversight should not be limited to initial sign-off but integrated into continuous postmarket performance evaluation.

For devices where the reference standard outcome that the AI was developed against cannot be collected automatically or can only be collected automatically through costly or burdensome processes, periodic calibration assessment of AI performance against its benchmark when approved is critical, as several factors may change over time and impact performance. Examples include if a sensor is sourced from a different company and its outputs are different, if material components of the device are changed causing the AI to require recalibration, or if the context of use changes causing sub-population performance to change. For AI that uses routinely collected information, data drift, clinical practice, and data collection processes can change and impact the AI.

Question 3. Postmarket Data Sources and Quality Management

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

Electronic health records, device logs, and patient-reported outcomes are all important sources for ongoing performance evaluation, as well as imaging DICOM (Digital Imaging and Communications in Medicine) data.

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

To address data quality and completeness, assessments of bias in sub-populations, documentation of data collection, and data coverage in multiple environments and settings are all performed.

For interoperability, standards are critical both for storing data at rest and for data in transit. Lastly, knowledge representation using controlled vocabularies and ontologies where possible is important to ensure computability of data and aggregation across sites and settings. A well-known example of a controlled vocabulary is LOINC. Well-known examples of data models for data at rest include OMOP and DICOM. Well-known examples for data in transit include FHIR and HL7.

Internal data validation procedures are executed within the ETL processes. During the evaluation, data quality assessments are performed with clinical or subject matter experts to identify any data aberrations that could affect the analysis. A procedure should be proactively developed for handling missing data, such as performing chart review or statistical imputations.

Question 4. Monitoring Triggers and Response Protocols

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

Surveillance of the AI's input data is conducted to ensure that sensor data and routinely collected data remain consistent with expected data distributions and density within established operating parameters, as deviations may indicate the need for further review and more in-depth evaluation. Additional evaluation may also be prompted by monitoring the AI's outputs for clinical appropriateness, manual guardrails, and data-driven outlier detection. Assessment of change in the context of use is conducted to determine if the device is being used outside of the clinical context in which it was tested and approved, which can trigger the need for additional evaluation.

Question 5. Human-AI Interaction and User Experience

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

Beyond formal guideline revisions, changes in workflow, clinician behavior, or patient populations can meaningfully affect AI device performance. Monitoring user interaction logs, override rates, and feedback can reveal early signs of drift in clinical use. Major updates to practice standards or task delegation should automatically trigger a device review and potentially a model update.