# Measuring and Evaluating AI-enabled Medical Device Performance in the Real-World

## Meta information

**Related docket**: FDA-2025-N-4203

**Submitter**: Prof. Dr. Christian Johner, Johner Institute, Author of the AI Questionnaire of EU TeamNB, christian.johner@johner-institut.de

**Date**: 2025-10-07

**Offer**: Please don't hesitate to contact submitter if there are any questions or need for help.

## Requested Comments

### 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?**

*There is no one-fits-all metric. We rather must encourage manufacturers to make sure that*

- *model parameters are aligned with product parameters / specifications and*
- *product parameters / specifications with intended use.*

*We observe that this traceability frequently is not given.*

*Furthermore, manufacturers should not only specify the performance metrics (to be) achieved but also specify the related confidence levels.*

*Typically, manufacturers specify metrics such as specific, sensitivity, AUC, etc. But these metrics are only suitable for selected classification tasks. For other tasks other metrics are more suitable such as*

- ***IoU (Intersection over Union)*** *= $|A \cap B| / |A \cup B|$: Proportion of correct predictions at an IoU threshold of 0.5. A threshold of 0.5 means that a prediction is considered a "hit" if it overlaps with the ground truth by ≥50%.*
  *Practical example: Model finds 10 suspicious regions, 9 of which have IoU ≥0.5 with actual lesions, 1 is a false positive, Precision@IoU=0.5 = 9/10 = 0.90*
- ***"Sørensen-Dice index"****: Measure of the overlap between predicted and actual segments (0: no overlap, 1: perfect overlap)*

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

*I don't have a comment.*

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

*The timeframe is infinite. Manufacturers targeting the EU market should keep in mind the AI Act requirements for generating and analyzing "protocols".*

*In other words: the evaluation is part of the post-market surveillance and therefore a proactive and continuous (infinite) activity.*

## 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?***

*We suggest four types of monitoring:*

1. *Monitoring if the **models' input** data e.g., whether patient population is still matching population specified in intended use, whether there are new types of device data or device settings or other parameters that might have an impact on the outputs.*
2. *Monitoring of the **model's output** e.g., the performance metrics, but also more specific analysis such as partial dependency plots, feature plots, depending on the models' architecture.*
3. *Monitoring of the **user interaction** with the model/device e.g., whether there are signs of use errors including over trust and mistrust.*
4. *Monitoring of the **technical environment** such as hardware, operating system, libraries, systems that the device is interacting with (e.g., to detect interoperability issues)*

**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?**

*We consider humans to be indispensable for the following activities:*

- *Ensuring the alignment of intended use and model performance*
- *Defining metrics and methods for evaluating the model performance*
- *Automating the evaluation*
- *Assessing the evaluation results*

*The automated monitoring helps with reproducible and repeated evaluation.*

*We don't see a need for balance. Both is required.*

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

We suggest:

- Generation of audit logs / protocols providing information about
    - Input data including data shifts
    - Output data including biases
    - User interaction
    - Technical environment
- Infrastructure to capture, collect and analyze this data
- Sandboxes to simulate changes in the model environment such as input data, runtime environment (e.g., hardware, operating system, libraries)

## 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)?**

*I do not understand the difference to the previous questions.*

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

*See above*

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

*We consider to be effective:*

- *Methods of **human factors engineering** such as observations, usability tests (summative evaluations), interviews, and analysis of user behavior (e.g., via 'protocols')*
- ***Comparison of target and actual population** using all attributes, that might haven an influence on model performance*
- ***Comparison of real-world model outputs / recommendations and clinical actions with clinical best practices** e.g., proposed by medical professionals or scientific literature. It is important, that the medical professionals are neither influenced / biased by the models' outputs / recommendations nor by other medical professionals. It turns out, that frequently the intra-person and inter-person variability of ground-truth-definition is higher than the deviation from model outputs from the best-practices.*

## Monitoring Triggers and Response Protocols

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

*Triggers are:*

- *More human factor problems than anticipated*
- *Deviation between target and actual population*
- *Changes to the technical environment*
- *Model and software updates including AI/ML libraries and LLMs (if used)*
- *New or updated hardware*
- *Negative customer feedback or clinical outcome*

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

*This depends on the metrics. For examples deviations can be defined by t-tests statistics.*

## 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?**

We observe: Higher workload, higher cognitive demand, shift of tasks (e.g., only most difficult decisions, less patient contact) with resulting fatigue, errors, and over-trust (just confirming the defaults).

More details can be found in our paper *"Usability Engineering for Medical Devices using Artificial Intelligence and Machine Learning Technology"*.

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

We compiled a comprehensive list of best practices in our AI guideline for medical devices.

For example, we recommend:

- Provide explanations to users why the AI reached a certain conclusion, without overwhelming the users with information.
- Don't let the engineers (e.g., software developers, data scientists) define the metrics, but rather the medical / clinical affairs teams.
- Don't limit the evaluation of the model performance to metrics, but additionally apply methods of AI interpretability to understand the model as for example described here.

## 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?**

All AI related pre- and post-market activities must be baked into the quality system. We frequently observe major gaps such as:

- The ground truth is confused with gold standard, and both are not commonly and precisely understood, documented / specified ("3 physicians, 4 opinions").
- Data labeling requirements are not defined specifically to the device; labeling experts are no real experts, fatigue is not considered,
- The clinical affairs team does not understand the technology and the state of the art but must judge the risk benefit ratio in the clinical evaluation.
- The AI specific risks are not fully identified, understood, and mitigated.
- The development process does not require to justify the selection of the model's architecture and hyperparameters, the different "attempts" are not documented, just the result. I.e. the "design history" is missing.

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

*I don't have a comment.*